The Use and Misuse of the Experimental Method in Social Psychology

A Critical Examination of Classical Research

AUGUSTINE BRANNIGAN

The Use and Misuse of the Experimental Method in Social Psychology

This book critically examines the work of a number of pioneers of social psychology, including legendary figures such as Kurt Lewin, Leon Festinger, Muzafer Sherif, Solomon Asch, Stanley Milgram, and Philip Zimbardo. Augustine Brannigan argues that the reliance of these psychologists on experimentation has led to questions around validity and replication of their studies.

The author explores new research and archival work relating to these studies and outlines a new approach to experimentation that repudiates the use of deception in human experiments and provides clues to how social psychology can re-articulate its premises and future lines of research. Based on the author's 2004 work *The Rise and Fall of Social Psychology,* in which he critiques the experimental methods used, the book advocates for a return to qualitative methods to redeem the essential social dimensions of social psychology.

Covering famous studies such as the Stanford Prison Experiment, Milgram's studies of obedience, Sherif's Robbers Cave, and Rosenhan's exposé of psychiatric institutions, this is essential and fascinating reading for students of social psychology, and the social sciences. It's also of interest to academics and researchers interested in engaging with a critical approach to classical social psychology, with a view to changing the future of this important discipline.

Augustine Brannigan is Professor Emeritus of Sociology at the University of Calgary, Canada. In his career as a professor in the Department of Sociology he taught social psychology, social theory, criminology, and criminal justice.

Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

The Use and Misuse of the Experimental Method in Social Psychology

A Critical Examination of Classical Research

Augustine Brannigan

Routledge
Taylor & Francis Group
LONDON AND NEW YORK

First published 2021
by Routledge
2 Park Square, Milton Park, Abingdon, Oxon OX14 4RN

and by Routledge
52 Vanderbilt Avenue, New York, NY 10017

Routledge is an imprint of the Taylor & Francis Group, an informa business

British Library Cataloguing-in-Publication Data
A catalogue record for this book is available from the British Library

Library of Congress Cataloging-in-Publication Data
A catalog record has been requested for this book

ISBN: 978-0-367-47310-5 (hbk)
ISBN: 978-0-367-45816-4 (pbk)
ISBN: 978-1-003-03480-3 (ebk)

Typeset in Times New Roman
by River Editorial Ltd, Devon, UK

To Terry Marilyn Brannigan, my lifelong companion

Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

Contents

Preface

The intellectual challenge of experimental social psychology in the classical period

This book is about the attempts over the past seventy or so years to forge a science of social life based on the systematic use of experiments. Experimental social psychology is unique in the social sciences in that it has committed itself, primarily in North America, almost exclusively to the use of the experimental method to create new knowledge. In my view, this attempt has wavered, and the rise of the institutional review boards and human ethics boards promises to bring the discipline founded on high-impact experiments based on deception of subjects to an end. The conclusion of this book is that experimental social psychology is, *at present*, an impossible science with little possibility, in its current configuration, of establishing any credible new knowledge. There are many reasons for this. The subject matter of the field is already part of the competence of people in everyday life. Certainly, scholars in other fields must face this problem, but social psychology occupies an unusual scientific space. Where the historian researches documents that capture, say, acts of genocide, the historian is merely telling a story that links the documents in a coherent way. His or her knowledge may be more "thorough" than the witnesses to genocide might individually report. By contrast, the experimental social psychologist claims to be exposing processes that explain genocide or other topics in a non-obvious way, typically with reference to processes and mechanisms of which the original actors are unaware. Otherwise, he or she is merely repeating the obvious without the advantage of the historian, whose research of the *primary* documents puts facts before the reader that might not be known otherwise.

Furthermore, the experimentalist has to conjure up a proxy, or a shorthand artifice or substitute, for the original event. Rather than going to primary sources to study the phenomenon first-hand, the experimentalist has to visualize a way of reducing the process to something that can be studied in a laboratory over a short period of time, whether or not this is the best method of elucidating the phenomenon. The result is not a study of genocide but a metaphor of genocide, a dramatization or allegory that enacts certain key processes that the psychologist feels are critical, though these are frequently researched in a complete empirical vacuum with respect to the original events that characterized the genocide.

In this process the experimentalist is at elevated risks of importing into the study deeper moral and/or philosophical presuppositions. In other words, not being constrained by any set of "hard facts" that arise from studious observations of the phenomenon *in situ* – what Fran Cherry called "the stubborn particulars" of everyday life – and not being informed by what is found in the historian's documents, or the clinician's interviews, or the demographer's age–gender tables, the moral substructure of social science inquiry is given free play. The evidence that I will examine in this book suggests that moral issues often make "consumers" of experimental social psychological research, students and the public at large, overlook the obvious empirical deficiencies of experimental designs. This explains the popularity of the field, and its attraction to psychology and sociology majors in spite of its scientific weaknesses. Social psychology is like divinity in the nineteenth-century liberal arts curriculum – interesting, but not really practical, deeply relevant to everyday life without being a source of definitive or scientific understanding of the social world.

The future of social psychology

This work explores these ideas in the case of a number of the classical contributions to experimental social psychology, experiments to which every student in the field would be exposed. I refer to the group influence tradition of Muzafer Sherif, Solomon Asch, Stanley Milgram, and Philip Zimbardo, all of whom were wedded to the experimental tradition, were seized by the problems of everyday life, and each of whom had a moral interest that animated his research and that explains its enduring appeal. This moral appeal is highly evident in the study of IQ and teacher expectations, and worker productivity and employment conditions in two classical studies of expectation effects: Pygmalion and Hawthorne. Both were methodologically flawed investigations that nonetheless captured the hearts and minds of millions of students and members of the public. Their attraction was in their moral subtext, not their findings. The same conclusions may be drawn from the studies of indifferent bystanders to violence, persons whose failure of altruism motivated studies to mimic the same accounts in the lab. These studies were initiated following the notorious rape and murder of "Kitty" Genovese, a case that has been revised in light of new information. The practical application of psychology is explored at length in the more recent experimental study of violent media and the allegedly worrisome effects of aggressive fiction on viewers. The rise of pornography in recent decades has fueled enormous public concerns about violence against women, and psychologists have enjoyed tremendous opportunities as expert witnesses in prosecutions trying to control sexually provocative fiction. I review the psychological evidence as well as its impact on the common law of obscenity to determine whether any "hard facts" were identified in this research and whether its treatment by the courts was warranted. I also explore whether there is any such thing as "social learning" and whether it explains the origins of antisocial behavior.

The balance of the book examines the impact of feminist and Darwinian agendas on the study of gender, although in this case, the methodologies are not always experimental. Again I ask whether the social relevance of the research is based on scientific considerations, or moral and philosophical perspectives parading as science, that is, “living better” through science.

The final chapters attempt to give an overall assessment of the achievements of the classical traditions which resulted in an epistemological crisis about the utility of experimental studies of social life based on deception (Chapter 9). We compare this crisis to the contemporary replication crisis in social psychology (Chapter 10). And we end by discussing the prospect of empirical studies of social interaction based on non-experimental methods. This book presents for the first time a useful overview of the recent archival studies of the classical experiments. In addition, it highlights the importance of non-experimental approaches found in various forms of discourse analysis which increasingly rival the importance of experimentation in forging the future of social psychology.

Acknowledgements

I want to thank the many friends and colleagues who read the manuscript and offered support and insight: Sharon L. Williams, Mike Lynch, Travis Hirschi, Michael Overington, Gwynne Nettler, Rod Cooper, Tullio Caputo, Rick Linden, Doug Skoog, Hank Stam, John Mueller, John Ellard, Erin Van Brunschot, J. Edwin Boyd, Bob Stebbins, Dick Wanner, Nick Jones, Ron Roisen, Nestar Russell, Gina Perry, Stephen Gibson, Matt Hollander, Jason Turowetz, Zhiqiu Lin, Alex Bierman, Allan Mackay, Thibault Le Texier, Rich Griggs, and Jim Frideres.

1 The sunset on a golden age

Reflections on the gap between promise and practice

Human experimentation and causal explanations: promise versus performance

In his recollections of the history of social psychology, Albert Pepitone (1999:182) noted that, during the golden period, theory-driven research developed “the state-of-the-art-experiment in which the theoretical variables were created through artful manipulations of how subjects should interpret the situation.” Such stagings often entailed “fantastic scenarios and ‘cover stories’ that stretched credulity … From that initial postwar boom … into the late 1980s, marks a golden age … The golden age was quintessentially the age of experimentalism.” The experiments were designed to formally test hypotheses about behavior. But such “sharply focused experiments cannot themselves acquire comprehensive knowledge” (1999:182). In fact, it produced a crisis in confidence about just how far the experimental tradition was capable of contributing to the growth of knowledge. According to Zimbardo, that period drew to a close when persons with less appetite for extreme dramaturgy joined the profession and when institutions took meaningful steps to safeguard the treatment of human subjects. “That tradition is now dead and not mourned by those who hastened its demise, a cabal of some cognitive social psychologists, human subjects research committees, Protestants, and female social psychologists” (Zimbardo 1999).

In my view, there was a huge gulf between the appropriate use of experiments and what psychologists actually did in their laboratories, and this was the case not at the periphery of social psychology, but at its very core. In other words, there was a gap between the promise and the performance after experimentation dominated the arsenal of social psychologists in the 1950s and subsequent decades. In this chapter, I want to identify the scientific strength of experimentation and contrast this to how experiments actually developed in the golden age of experimental social psychology.

The promise

In the methodology of the social sciences, it is well accepted that experimentation is the key to objective knowledge, and is superior to rival methodologies, at least in principle.[1] The ideal design in scientific research is the *true* experiment, where subjects are randomly exposed to various treatment conditions and then tested to determine the effects of the different treatments on the outcomes. Since the designs are standardized, replication of results is typically quite straightforward. What made the experiment superior to other methods, such as cross-sectional surveys, ethnographies, or interviews, was, according to its proponents, its ability to combine certain features of inquiry: first, an association between two or more variables linking a potential cause and an effect; second, the ability to identify temporal precedence of the cause, that is, the appearance of the cause *prior* to the identification of the effect because of the time-ordering of the events in the experimental designs; and third, an ability to determine whether the connection between cause and effect is non-spurious. In addition, true experiments have three things associated with them: two comparison groups – minimally, an experimental group and a control group, variations in the independent variable before assessment of change in the dependent variable, and random assignment of subjects to the two (or more) groups.

In theory, this combination of factors is supposed to give some confidence in the validity of the causal connections between the "treatment" and the outcomes. And our confidence is further enhanced by two things: the identification of the causal *mechanisms* that underlie the observed changes and the experimenter's control over the institutional context of the experiment. This said, it must also be acknowledged that not all experiments have a pure "exposure" and control or "no exposure" design. Sometimes a design will have several different *kinds* of exposures. Imagine looking at the effect of exposure to violent films, versus non-violent films, versus no films at all. The "no film" condition would be the true control group and the other types of exposures would be comparison groups across experimental treatments. In addition, true experiments do not actually require *pre-tests* on the variable or outcome of interest. If one was interested in the effects of certain types of films on attitudes (for example, propaganda and attitudes to certain minorities), it might be possible to get a pre-test measure of attitudes *prior* to the treatment. However, the logic of the design is that those in the control group are, in effect, a pre-test group because they have not received the treatment. Because they have been randomly assigned to the control condition, they are logically identical to the "before" group. The random assignment of subjects to various treatment groups avoids the potential artefact that arises from administering the same measures to the same subjects twice.

A related point has to do with randomization. Of course, it would be a mistake to believe that the people who end up in the treatment versus the control group are all exactly alike. They obviously are not. But what the logic of random assignment suggests is that the various salient things that might affect the outcomes have an equal probability of occurring in each group, so that their effect is neutral.

A different issue concerns random *sampling* versus random assignment. In a survey, we engage in a random sampling of a population to ensure that we can generalize from the persons in the sample to the larger population, since each *S* has the same probability of being selected for inclusion in the survey (Frankford-Nachmias 1999:481). By contrast, in experiments, randomization does not ensure generalizability, and neither is it designed to do so. It is designed to ensure internal validity. Internal validity covers a number of issues, but, for our purposes, suffice it to say that what it ensures is that the design gives us confidence that the only important difference between the control and the treatment group is the treatment itself. The issue of generalizability is perhaps the main Achilles' heel of experimentation in social psychology. The logic of experimentation is that the sorts of things being investigated are of such generality that they are present in whatever sections of the population from which subjects are drawn. At least in theory, the lack of a careful selection process designed to ensure representativeness is irrelevant. Obviously, with these attributes, the experiment has earned a reputation as a powerful tool in the arsenal of social scientists. How did the experiment work out in practice? A rather different picture emerges.

The golden age of experimental social psychology: reflections from the Yosemite conference

In 1997, a group of senior American social psychologists gathered at Yosemite National Park to take stock of the growth of knowledge in experimental social psychology and to record some of their personal memories and professional reflections. They were the leading lights in psychology, who were active in creating the profession in the period of its heyday following the Second World War and in the decades thereafter.[2] All the participants took their doctoral training in the period 1948–1959, and were major contributors to the field during its impressive growth in the 1950s, 1960s, 1970s, and 1980s. Their post-mortem deliberations on the achievements of the field provide a rare window on the history of social psychology. They identify a number of ambiguities in the development of the field that appears to be associated with its commitment to the causal explanation of social behavior through the use of experimental methods.

What important conclusions emerged from this celebration of a century of research? There were recurrent observations that the golden age was over, that the field had not accumulated much reliable new knowledge, and that it had not achieved much consensus about important matters. The sociology of knowledge warns us to treat collective reflections about a field's origins with a grain of salt, since often times these "memories" are *myths* about the turning points in history that eventuated in the current configuration of knowledge. Rodrigues and Levine (1999), who edited the proceedings, trace experimental social psychology to the work of Norman Triplett. In 1897, Triplett had published an experimental study of children's task performance based on observations of bicycle racers. Triplett knew that the individual performance levels of racers are influenced by competition. Typically, racers "pace" each other before putting on the final sprint at the end of the race. Triplett measured children's performance on a fishing reel undertaken alone and in competition. His work introduced the concept of "social facilitation" into individual productivity as well as the effects of the related concept of "rivalry." What is ironic is that his research received rather mixed notice in the years that followed, and, aside from being experimental, created no research legacy.[3] Indeed, in the early decades of the discipline, G. H. Mead criticized the use of experiments for social psychology, since he favored the method of introspection. However, social psychology gradually acquired scientific respectability, not because of its theoretical progress, but because of the adoption of the methods and logic of the hard sciences. The break between social psychology as a *sociological* discipline and social psychology as a *psychological* discipline emerged in the 1950s, when experimental methodology became the orthodox approach in professional psychology while the sister sciences remained relatively diversified in their approaches. Experimentation established itself as the gold standard because of its ability to link connections in a non-spurious, temporally informed fashion, and to explore relationships causally.

Several contributors at the Yosemite conference noted that social psychology had become a field in which practitioners appeared to know little of the history of their own discipline and had become alienated from cognate areas in sociology and anthropology. In this celebration of the discipline's achievements, Aronson (1999:108) lamented the fact that contemporary social psychologists were ignorant of research prior to 1975, and Raven (1999:118) warned that new scholars were in danger of "reinventing the wheel," or of failing to credit an idea to its originator because of disciplinary amnesia (Berkowitz 1999:161). There was a consensus that the field had become increasingly abstract, specialized, and divorced from issues of everyday life. There was also a sense that "the golden age" of experimentation had come to an end, a victim of the new institutional review boards instituted in the 1990s to protect human subjects from unethical conduct by experimenters.

The review boards were created in the 1970s to ensure protection of human subjects from harm and discomfort.[4] While originally directed at medical research using human subjects, the boards may have sounded the death-knell for experimental studies of human psychology. In such studies, consent is often obtained from subjects through deception about the purpose of the research, a condition that renders the consent uninformed, and, hence, invalid. In addition, in the search for realism in the laboratory, some psychological experiments have entailed very detailed dramaturgical manipulations that have resulted in high levels of trauma among subjects. For example, in the disturbing study of obedience to authority, Milgram (1963) reported that many of his subjects experienced nervous fits and uncontrollable seizures. Zillmann and Bryant's (1982) nine-week study of pornography reported that changes in callous attitudes among subjects were "non-transitory" (i.e., permanent). In their study of the dynamics of emotions, Schachter and Singer (1962) injected subjects with chlorpromazine (a medication used in the treatment of schizophrenia) or epinephrine (synthetic adrenaline) under the pretext of testing a new vitamin, "suproxin" (Schachter 1971). The passing of high-impact experimentation was noted by Zimbardo, mentioned earlier in this chapter. Zimbardo claimed that the ethics review boards "overreacted to the questionable ethics of some of the research by the oldies but goodies in experimental social psychology." By imposing limits on what can now be done and said to research participants, the boards have provided safeguards "to the end of eliminating some of what could be called traditional experimental social psychology" (1999:138).

The dominant understanding about the defensible treatment of human subjects in experimental psychology has been that there has existed a trade-off between short-term deception and edgy manipulation of subjects on the one hand, and long-term benefits to science and society on the other. But there has never been a general meeting of the consumers of psychological knowledge to determine whether this investment was justified. In fact, Zimbardo's own work raises some of the deepest questions. In the Stanford "prison study," he reports that some of his mock guards assaulted the mock prisoners, and that many prisoners had to be released prematurely because of intense emotional trauma. In his 1972 account in *Society*, it appears he dragged his heels in terminating the "experiment" until it could be recorded on videotape by a local television station. But what was learned about prisons that we did not know? If the ethics boards overreacted in recent years, this may be due to an absence of effective internal self-regulation in the past. But that was not the only problem associated with experiments in the golden age. Evidence suggests that in many of the key studies, researchers would not take no for an answer, that the experiments simply became devices for demonstrating a relationship arrived at beforehand, and that the field could not grow because falsification of a hypothesis was virtually never recognized. In fact, "verification bias" has been identified as a leading cause of replication failures in contemporary experimental psychology.

Kurt Lewin and field theory

In my view, experimental social psychology began, not with Triplett in 1897, but with Kurt Lewin and his students in the mid- to late 1940s. Lewin was a German émigré whose "field theory" (1951) was based on the German gestalt tradition in which individual actions and attitudes were interpenetrated by socially based, cognitively coherent frames of reference. After the war, Lewin established the Research Center for Group Dynamics at the Massachusetts Institute of Technology. The Center moved to the University of Michigan after Lewin's death in 1947, but not before Lewin had assembled an impressive group of graduate students and co-instructors. By all accounts, Lewin was an effective "tribal leader" (according to Deutsch 1999:9). He combined a commitment to the rigors of experimentation with an intellectual agenda that fostered practical engagement with everyday life, including the potential of social psychology for ameliorating social problems. Lewin's own work demonstrated the greater effectiveness of democratic versus autocratic forms of leadership in laboratory studies. This in itself had a tremendous attraction among the graduate students who enrolled in social psychology after participating in the war against European and Japanese dictators. The field had a further caché since experimentation was the sole methodology in the social sciences expressly capable of suggesting not mere correlations, but causal connections. At the Yosemite meeting, Harold Kelley recorded the attitude at the time: “We were ‘real scientists,’ using the experimental method, drawing firm conclusions about cause and effect, and not fooling around with mushy correlational data” (1999:41). According to Gerard (1999:49) Cohen and Nagel’s *Introduction to Logic and Scientific Method* (1934) was standard reading for these scientific protégés.

The late 1940s and early 1950s were very consequential for the subsequent directions of the discipline. Lewin’s gestalt orientation appeared to lead psychologists away from a focus on stimulus–response behaviorism in their theoretical modelling, since it expressly celebrated the distinctive role of perception, recollection, and normative action in human social behavior. However, the commitment to the method of experimentation was subsequently to result in a narrow focus on cognitive *mechanisms* that were suitable for laboratory investigation, however remote they might be from pressing problems in everyday life. This narrow methodological focus was subsequently to stifle the search for general, integrative theories. By contrast, at the start of their careers, Deutsch, Gerard, Berkowitz, and Pepitone (among others) attested to the remarkable breadth of social psychology, and to the common definitions of problems in “sociological” and “psychological” social psychology. There was a tremendous interest in racism and discrimination, the related problems of school desegregation, and concerns over world peace. As the field developed, experimentalists became increasingly preoccupied with, in the words of Berkowitz (1999:162), “within-the-skin” versus “between-skins” phenomena. The “social” in social psychology became more associated with the idea of “information” and “information processing” than with meaning, culture, and context. And the focus on “cognitions” *qua* cognitions created “a spurious conceptual generality” (Pepitone 1999:191).

One of the casualties in the institutionalization of social psychology was psychoanalysis. Morton Deutsch, one of the earliest proponents of experimental studies of group behavior, was trained in psychoanalysis and enjoyed a long clinical career outside academic psychology. “The practice was personally rewarding. I helped a number of people, it enabled me to stay in touch with my own inner life, and it provided a welcome supplement to my academic salary” (Deutsch 1999:15). Similarly, Harold Gerard turned to “depth psychology” later in his career, entering psychoanalysis at age fifty-nine and becoming a psychoanalyst at age sixty-nine – thereafter switching his experimental work to a focus on “subliminal activation.” But these clearly were the exceptions.

As the field became more laboratory-oriented, such “soft methods” and general theories fell into disfavor. The single leading proponent of experimental social psychology – Leon Festinger – viewed such “applications” with suspicion. According to Aronson, “Leon was not interested in improving the human condition. Not in the least … Trying to understand human nature and doing good research (not doing good) were more than enough to keep him excited" (1999:87). Indeed, Festinger held Aronson's other mentor, Abram Maslow, a leading humanist, in obvious contempt. Festinger once told Aronson, "That guy's ideas are so bad that they aren't even wrong" (1999:92). In the end, Festinger's scientific model overshadowed the humanistic methods and theories of Maslow, and Freud's theories became unwelcome in the era of experimentation.

The impact of cognitive dissonance

Festinger's attitude to humanism is difficult to fathom. In his own work, he attempted to ground research in provocative issues taken from everyday life. The dynamics of everyday life could be distilled in the formal light of causal ordering in the controlled experiment and applied back to the outside world. His study of group conformity and rejection of deviants arose from a field study of political consciousness among graduate students in college residence (Festinger, Schachter and Back 1950). The famous field study, *When Prophecy Fails* (Festinger, Riecken and Schachter 1956), was one of the most important intellectual foundations for his studies of cognitive dissonance. In *When Prophecy Fails*, Festinger et al. discovered that a failed prediction of the imminent destruction of the earth led its proponents to become *more* attracted to the prophecy after its failure, not *less* so, as might be predicted by behaviorism. This study had far more potential for understanding irrational behavior than anything subsequently "discovered" empirically in the laboratory. The passing of such brilliant fieldwork was another casualty of the experimental "turn" in psychological methods. This brings me to what I believe was the pivotal role of the experimental study of cognitive dissonance, a move that appears at several levels to be indicative of the sea change that occurred in social psychology.

Aronson reported at Yosemite that he read Festinger's manuscript, *A Theory of Cognitive Dissonance*, in typescript as a graduate student at Stanford, and that he found it revolutionary:

> [It] revitalized social psychology … and offered a serious vehicle for challenging the smug dominance of reinforcement theory. It did this not in a vague, philosophical manner, but in a powerful, concrete, and specific confrontation, exposing reinforcement theory's limiting conditions as well as its inability to predict some of the more subtle and more interesting nuances of human behavior.
>
> (1999:86)

This was strong testimony to the impact of Festinger's vision on the field, and the agenda-setting implications for his students in social psychology. When he arrived at Stanford, Aronson (1999:85) reported that he eagerly enrolled in what he referred to as Festinger's "seminary" in social psychology. What is ironic is that, prior to joining the faculty at the Massachusetts Institute of Technology (and later Stanford), Festinger had not taken a course in social psychology, and neither had his leading student, Stan Schachter (see Festinger 1987:2), nor, apparently, had another leading light in the field, Donald Campbell (see Raven 1999:115). Furthermore, those who had studied psychology at the New York universities as undergraduates experienced little indoctrination in behaviorism (if the work of Solomon Asch or Floyd Allport is any guide). The new generation of social psychologists appeared unacquainted with the leading social psychology texts of the 1930s, which, in any case, were *not* primarily behaviorist. Cognitive dissonance simply opened a new page for social psychology, one that was dogmatically experimental, indifferent to humanitarian agendas, and focused on the underlying mechanisms of sense-making.

In debates about psychology's apparent lack of progress, reference is often made to the "infancy" of the field. I believe another hypothesis should be examined: the potential for growth in social psychology perhaps was stifled precisely *because of* the impact of cognitive dissonance and the subsequent "cognitive revolution" in the field, something fostered by the dominance of the experimental method. The concept of cognitive dissonance is based on the idea that the human mind resists the simultaneous appropriation of two cognitions that are inconsistent. If one believed dearly that the world was going to end, and shed all one's worldly goods in anticipation, the realization that the world had *not* ended when expected would create a cognitive disequilibrium that would create a drive designed to restore cognitive consonance. One would have to accept the fact that one's beliefs were mistaken, and that one had foolishly given away one's material resources. In that way, all the facts would cohere consistently. Alternatively, one could reject the failure of the prophecy and revise the date of doomsday. Either way, the mind seeks an equilibrium in "cognitions." This is what Festinger had recognized conceptually before taking the matter to the laboratory.

In the laboratory, Festinger and Carlsmith (1959) enlisted naïve subjects in a series of repetitive, ostensibly boring, mechanical exercises. The naïve subjects were told that the experimenter's assistant was unable to instruct new subjects, and that a replacement was required. Subjects were recruited to replace the assistant, and to coach new subjects. Specifically, the recruits were required to tell the new subjects that the experiment was very exciting and interesting when nothing could be further from the truth. After the naïve subjects (turned recruits) had delivered this false information (to a confederate of the experimenter), they were tested to determine what they themselves thought about the experiment. They were asked *how enjoyable* the experiment was, how much they thought they had *learned,* how scientifically *important* it was, and their *desire to participate* in similar experiments. The first critical dimension of the experiment was that the subjects were manipulated into lying to someone they thought was as naïve as they themselves had been. The second critical dimension was the amount of money they were paid as recruits to (mis)inform the new subjects (who, it turns out, were confederates). In the one case, they were paid a large amount of money (twenty dollars), in the other they were paid a small amount of money (one dollar).

What was the effect of the level of compensation on the subjects' personal estimation of the degree to which the experiment was interesting, exciting, etc.? If you were paid a lot of money to lie about such matters (twenty dollars), would this salve your conscience? Would this payment mollify the lie, and leave your estimation of the experiences unaltered? On the contrary, if you lied for peanuts, would you experience an anomaly that would compel your mind to think that the experiment was more exciting in retrospect than it had been originally, and that, as a consequence, you had not really lied at all? In short, did the dissonance (between what you experienced personally and what you told the other) create processes in which your sense-making devices reordered the significance attached to the original events below the threshold of consciousness after you were paid? Or did the compensation (twenty dollars) in effect justify the misinformation given to others, and leave your original memories unaltered? If so, cognitive dissonance would become a paradigm for opinion consolidation and change, and, presumably, predict some concordance between opinions, values, and behaviors. *When Prophecy Fails* suggested this happened in everyday life. What was revealed when the experimental microscope was focused downwards to look more closely at the phenomenon in the laboratory?

The original experiment reported four predictions, but only one was corroborated (i.e., estimation of "enjoyment"). Persons paid poorly found the experiment more enjoyable than persons paid well. The cup was three-quarters empty, but, rather than re-evaluate the theory, the experimentalists returned to the laboratory. There followed a torrent of tests of cognitive dissonance exploring the dynamics of dissonance in myriad innovative, experimental contexts.

In 1964, Chapanis and Chapanis summarized the findings from several dozen tests of the theory of cognitive dissonance. They concluded that there was no evidence that subjects actually experienced the dissonance that the experiments were attempting to induce, that the manipulations were not credible social situations (paying someone a day's salary – twenty dollars – for half an hour's work – 16% voiced suspicions or refused to be hired in the original Festinger and Carlsmith design), that the designs were confounded by classical reinforcement processes, that there were massive problems in the rejection of cases and arbitrary reallocation across treatment groups, and that some studies used unusually permissive statistical criteria for rejecting the null hypothesis. "Experimental manipulations are usually so complex and the crucial variables so confounded that no valid conclusions could be drawn from the data … A number of fundamental methodological inadequacies in the analysis of results … vitiate the results" (Chapanis and Chapanis 1964:1). They reiterated Asch's (1958) conclusion in his review of Festinger's book: the case for cognitive dissonance was "not proven" (Chapanis and Chapanis 1964:22).

The disappearance of falsification and the decline in consensus

One would have thought that such damaging criticisms would have given serious pause to the field. But that does not appear to be the case. In *Reflections* *on 100 Years* (Rodrigues and Levine 1999), there are dozens of references to dissonance theory but not a single reference to Chapanis and Chapanis. Chapanis and Chapanis were not part of Lewin's tribe, and those who were simply ignored the critique! Not surprisingly, the report of negative findings was quite rare in experimental social psychology during this period, as noted by Pepitone: "An overall summary of experimental research would show that findings infrequently led to an outright rejection of the hypothesis being tested … in the vast majority of published research articles findings confirm hypotheses" (1999:193).

The problem with this is that a science that cannot say no to anything does not actually have the capacity to grow. Experiments had taken on a life of their own, and research had lost contact with everyday life. Events researched in one laboratory were designed to explore effects not found in everyday life, but in other laboratories. Zajonc (1997:200–1) voiced some difficulties with the consequences in terms of the progress of the field. He noted, looking backwards, that "social psychology (like psychology itself) is not cumulative." If one were to take any textbook, he says, and randomly reshuffle the chapters, it wouldn't matter, since "there is no *compelling* order." So, in a century of psychology, nothing accumulates. As we shall see, this observation has been recurrent in the history of the discipline. Zajonc then noted that "the scientists of a given discipline agree about the core subject matter of their inquiry … But psychologists and social psychologists do not. We have no consensus about the core of our field's subject matter." Zajonc attributed this to disagreement about the fundamental characteristics of human nature (whether the mind is rational or irrational), but I would say that this cannot logically be a *cause* of the lack of consensus as much as another manifestation of it. In his reflections, Pepitone similarly noted that despite the volume of brilliant and creative work produced by experimentalists, "the theory–research programs have produced few absolutely general, context free, and universally valid principles or laws" (1999:192). Pepitone further observed that research traditions in social psychology seem to come and go like fashions, without achieving higher-order conceptual integration, or acquiring any enduring set of "hard facts."

In the 1970s, there was widespread discussion about the lack of progress, relevance, and consensus in psychology. This is sometimes referred to as the "crisis literature." At the time, many critics pointed to the liability of the experiment whose scientific allure outstripped its actual potential, since the majority of the experiments involved low-impact, short-duration interactions between strangers drawn overwhelmingly from the ranks of college sophomores. High-impact designs such as that employed by Milgram eventually raised ethical questions for members of the ethics review boards. In their textbook, *Experimental Social Psychology*, Murphy, Murphy, and Newcomb had warned of the limitations of the experimental method and its potential misapplication decades before experimentation became the *sine qua non* of social psychology: "It has become very evident in recent years that the social psychologist … has succeeded in experimental and quantitative control by leaving out most of the variables about which we really need to know" (1937:10). It is not surprising that

Albert Pepitone came to the same conclusion when he suggested that the apparent lack of progress derives from the *"successful* development of the lab experiment as the principal method of testing hypotheses and the principal source of hypotheses." As for the lack of theoretical progress, "This deficiency is also due to the exclusive use of experiments" (1999:193). Levine and Rodrigues appeared to agree: "Many of the current criticisms of the field today – for example, an overemphasis on experimentation, a lack of humanism, an unwillingness to focus on the applied – are part of Festinger's legacy" (1999:218). Festinger himself appears to have abandoned experimental social psychology in the mid-1960s after cognitive dissonance encountered heavy weather. But the field persisted and continued to examine social processes in the lab at an ethereal level, testing for relationships in a pure or context-free fashion, looking for what might be called a geometry of interaction, divorced from the bite of infernal life. Gerard reported that

> by the 1970s, social psychology had become dominated by the cognitive revolution that had swept most of psychology … I developed a sinking feeling that we social psychologists were missing the boat … I became dissatisfied with the bland cast that had overtaken social psychology.
>
> (1999:67)

Pepitone similarly observed that the experiment "systematically constrains the field to leave out of theory and research much of what is observed about the influence of culture and social structure" (1999:193). Thus, the approach that gave the field its scientific credibility constrained how problems were defined and actually inhibited its growth.

Situational analysis and the experiment

Another bias that students of Festinger and Lewin seem to have inherited is a sense of the primacy of "the situation" as a fundamental fact of social life. This is a legacy of the phenomenology that underpins gestalt psychology. Phenomenology emphasizes the role of embodiment, consciousness, and temporality in our immediate experience of reality. But these may not be important from the point of view of causal explanation. Borrowing from another context, this is evident in Jack Katz's criminology, in which he attempted to explain participation in crime by reference to the foreground of experience: the embodied attractions of doing evil. The problem he ran into was that things like robbery, no matter how pleasurable and seductive, had certain recurrent features: robbery is overwhelmingly the pursuit of youth, of males, and of poor blacks, characteristics that point to background structures. Phenomenology can describe processes, but an explanation that equates exogenous causes with an account of the processes that need explaining is circular.

The parallel deficiency in social psychology is the importance that is attached to "the situation." The experiment is premised on the idea that the essential elements in social interaction are situational and can be scripted into short-term experimental dramas. This overlooks compelling evidence from life-course studies about the stability of traits such as temperament and aggressiveness over the life cycle (Glueck and Glueck 1950; Moffitt 1993; Mischel 2014).

A good illustration of the frailty of this line of thinking was suggested by Aronson's discussion of anti-Semitism. At the Yosemite conference, Aronson (1999:104) reported that he was forged as a social psychologist at the age of nine following harassment by neighborhood kids who intimidated him en route to Hebrew school. "They were caught up in a powerful situation that produced that prejudice" (1999:104). As a social psychologist, he said he strives to create interventions that can produce "redemption" through a situational remedy. In other words, both the causes and the remedy to anti-Semitism are "situational." This flies in the face of national and institutional patterns of animosity that have marked centuries of European history from the time of the Romans. This "situationism" is not the legacy of thinking experimentally, but the legacy of thinking *only* experimentally. What the Yosemite conference suggests is that in the pursuit of a methodology designed to confer confidence in causal connections, social psychology lost some of its purchase on the complexity of everyday life, on depth psychology and emotional attachments, on life-course persistent traits, and on much of what preoccupies us as requiring social reform. The rise of the institutional review boards that censure deceptive cover stories and threatening manipulations – the stuff of classical experimental social psychology – might have been a blessing in disguise.

Is social psychology an objective science?

Social psychology is paradoxical. It is one of the most popular subjects in the undergraduate curriculum but very little of its subject matter is practical, or can be applied in specific settings to overcome problems that would plague society in its absence. It strikes students as highly "relevant" but its theories do not exhibit evidence of accumulation either in empirical findings or in the consolidation of non-obvious theories. The scope of the field is colossal, but its achievements are questionable. It does not appear to have generated a set of "hard facts" or "main effects" that ground a consensus of what is truly important. What coherence does exist appears to derive from adoption of the method of Galileo – experimentation – arguably the most powerful technique for making causal inferences in the relationships between social variables. However, it is unclear whether the underlying subject matter, that is, human behavior, is law-like and can be studied in an objective, value-free science. In addition, many of the key experiments are allegorical, theatrical, or metaphoric, since the important questions that arise in the study of everyday life often are rarely amenable to short-term, low-impact situations, and neither can the important things that preoccupy us – violence, trauma, misery – actually be examined experimentally with human subjects for obvious ethical reasons. The rise of institutional review boards designed to mitigate the harm to human subjects in medical and social science research, as noted earlier, may spell the end of an approach to social research frequently premised on the deception of the subjects. The field shows evidence of painful crises of confidence as investigators confront the disappointments and contradictions of the field. The Yosemite conference provides ample evidence of this crisis, but misgivings about social psychology's progress had been common in the field for decades. We turn to some of that literature now.

Notes

1 For example, Campbell and Stanley, reflecting on McCall's pioneering work on the use of experiments in educational research, hold up the experiment

> as the only means for settling disputes regarding educational practice, as the only way of verifying educational improvements, and as the only way of establishing a cumulative tradition in which improvements can be introduced without the danger of a faddish discard of old wisdom in favor of inferior novelties.
>
> (1963:2)

2 The participants included Morton Deutsch, Harold H. Kelley, Harold B. Gerard, Elliot Aronson, Bertram H. Raven, Philip G. Zimbardo, Leonard Berkowitz, Albert Pepitone, and Robert Zajonc. Stanley Schachter's terminal illness unfortunately precluded his attendance.

3 Triplett's legacy might have been better served had he played a role in educating graduate students who would continue his work. In order to assess the impact of his work, I examined the collection of social psychology texts published prior to 1950. The University of Calgary collection holds twenty-one such volumes. Triplett is cited in five (Allport 1924; Murchison 1935; Murphy, Murphy and Newcomb 1937; Newcomb and Hartley 1947; LaPiere and Farnsworth 1949). There is no reference in the other sixteen volumes (Dewey [1922] 1950, 1901; Ross 1908; McDougall 1919; Znaniecki 1925; Robinson 1930; Karpf 1932; Perry 1935; Sherif 1936; Hopkins 1938; Ginsberg 1942; Lowy 1944; Klineberg 1948; Krech and Crutchfield 1948; Blum 1949; Lindesmith and Strauss 1949). While not dismissing his impact, one would have thought that work considered pioneering would have attracted more consistent attention.

4 In the US, the National Research Act of 1974 established the existence of IRBs to oversee biomedical and behavioral research. In the UK, the Department of Health recommended the creation of Research Ethics Committees as early as 1968 but did not formally delegate responsibility to local research ethics committees until 1991.

2 Crisis and controversy in classical social psychology

Crisis and controversy in classical social psychology: self-doubts in a causal science

In *The Story of Psychology*, Morton Hunt (1993) describes the situation of social psychology this way:

> What extremely busy and productive field of modern psychology has no clear-cut identity and not even a generally accepted definition? Social Psychology. It is less a field than a no-man's land between psychology and sociology, overlapping each and impinging on several other social sciences.

Is social psychology a "no-man's land" that does not have a commonly accepted definition? How could such an ill-defined field prove so popular as an undergraduate subject? What is its attraction? As we enter the third decade of the century it is essential to take stock of our scientific progress to determine whether the questions we have raised and the methods we have used to explore them have actually yielded genuinely new knowledge and have perceptibly moved the field forward. As an undergraduate student, I subscribed to a wide range of courses in the social sciences, including social psychology. Psychology struck me as extremely relevant to everyday life, and as a profoundly relevant way of understanding human nature and the trials and tribulations that mark our ordinary lives. I was also struck by psychology's commitment to a methodology believed to be superior to those found in the other social sciences, such as sociology and economics. Experimentation epitomized the scientific mentality. It permitted the psychologist to test his or her ideas under controlled settings in which causal inferences were valid, and in which it would be possible to distinguish direct, indirect, and interactional effects. On this basis, a genuine science of human nature would come within our grasp. And, with it, people could design social arrangements capable of mitigating harm and injustice and encouraging the open development of self-expression – in other words, *Walden Two* in post-industrial society, without imperilling freedom and dignity.[1]

However, even as an undergraduate I was impatient with the overreliance on knowledge derived from the study of laboratory rats in artificial conditions. The comparability of rats and humans was taken on faith. To be fair, the parallels were drawn between basic learning processes of reinforcement, and not higher cognitive functions. I was also uncomfortable with the confinement of methodology to experimentation and the exclusion of research questions that could not be tackled within that framework. And, finally, I looked without consolation for any persuasive perspective that integrated the field theoretically. Except for experimentation (methodology), academic psychology seemed fragmented. And what appeared the most interesting and speculative development in the century – Freud's study of the unconscious – was beyond the pale conceptually, methodologically, and clinically. Rat conditioning attracted more confidence than the interpretation of dreams.

Nonetheless, much of the substance of the field intrigued me, particularly what I have come to view as the classic contributions. Later in life, as a professor teaching courses in social psychology, my interest in the questions that motivated these classic studies has continued to grow, but so have my misgivings about their theoretical and methodological premises. This book attempts to lay out my concerns, and to do so in the context of work that will be familiar to most students of psychology.

It is hard to convey to an audience in Europe, Asia, or Africa the importance attached to experimentation in social psychology in North America. That should not be surprising, since it is hard to convey it within North America to academics who are not psychologists. It is the dominant methodological approach, and it has become one pursued to the virtual exclusion of every other methodological strategy in the social sciences. As Stam, Radtke, and Lubek argue, "The acceptability of ideas in the field came to depend largely on the ability of authors to couch them in the language of the experiment" (2000:365). Yet, it has been used to tackle some of the trickiest social situations known to humankind. These include the general formation of social norms in Sherif's experiment on the autokinetic effect, the causes of compliance to the European holocaust in Milgram's study of obedience, the nature of workplace determinants of productivity in the Hawthorne studies, the causes of minority school failure in Rosenthal's study of teacher expectations, the contributions of violent and pornographic media to aggression and misogyny in everyday life in the work of Bandura and Donnerstein, among others. These studies are just a sampling of the classic contributions.

My interest in the work of social psychologists has been deepened by my own studies of the causes of crime and delinquency and by the evidence amassed in criminology about patterns of crime and its analogues, the characteristics of offenders and their social distribution in society. Is racial–ethnic hate crime explained by obedience to authority? Is street crime explained by youthful exposure to media violence? Both propositions would seem obvious to the student of social psychology, but neither would attract much confidence from students of criminology. How could the fields be so mutually insular? When the leading criminology textbooks view the contribution of social learning from mass media models to imitative crime as minimal, are they ignoring sound science? Or do the experimental investigations of media effects so overstate the explained variance of media exposure as to overshadow far more important determinants of violence (gender, class, and individual impulsiveness)? Much hangs in the balance.

For example, a number of national inquiries in Canada, the United States, Australia, and elsewhere have sought input from social sciences in respect of the effects of violent media in order to set guidelines for fictional portrayals of violent and erotic stories. In some jurisdictions, the courts have heard evidence from experimental social psychologists to determine if certain films are so threatening to public security as to put the liberty and capital of individuals who distribute them at risk under criminal obscenity laws. In fact, the introduction of work from experimental social psychology has been important in many areas of litigation, including forced busing in the United States, limits on industrial action at the Chancery Court in the United Kingdom, and the suppression of paedophilia and pornography in the United States and Canada. For many psychologists, acknowledgment of their accomplishments in legal decisions is strong corroboration of their social relevance and a vindication of their methodology. I cannot think of comparable contributions to jurisprudence from sociologists or criminologists, and I believe that the advantage that psychologists claim is their adherence to a superior methodology.

But the relevance of this controversy is not limited to the field of social psychology and public policy. There is a lively debate about the epistemology of social science in what has come to be called the "science wars." What is the evidence for the "existential determination" of the content of science?[2] The Sokal hoax was a telling chapter in this debate. Alan Sokal, a New York physicist who taught mathematics to peasant children during the Sandinista revolution in Nicaragua, duped the editors of *Social Text* with a bogus description of what he labeled a poststructural physics. His "transformative hermeneutics of quantum gravity" was a spoof on the claims about science running through the leading figures of the French intellectual establishment spiked with the usual impenetrable jargon and buzzwords (Sokal 1996a). The parody was published in a special issue of *Social Text* designed by the editors to rebuke the criticisms of their antiscientific agenda by leading scientists. Sokal simultaneously published an exposé in *Lingua Franca* (Sokal 1996b).

In a more recent study, Simmons, Nelson, and Simonsohn (2011:1360) showed that exposing subjects to different children's songs induces an age contrast "making people feel older." In a related study, listening to the Beatles' "When I'm Sixty-Four" versus a control song made them feel younger. Their point was "undisclosed flexibility in data collection and analysis allows presenting anything as significant." Researchers have enormous degrees of freedom in choosing among dependent variables, choosing sample size, employing alternative covariates and reporting subsets of experimental conditions – making it extremely easy for experimentalists to fail to reject the null hypothesis based on a 0.05 probability level. The lesson is that this happens all too frequently in experimental work. My current inquiry raises the question of the extra-scientific determination of conclusions in the area of psychological knowledge, and the autonomy of social scientific knowledge from everyday life. If my suspicions are borne out, much of what passes for science in psychology is morality in an experimental idiom.

The topic of experimentation merits re-evaluation on other grounds. There is recent historiography in respect of one classic study – Milgram's study *Obedience to Authority* (1974) – which suggests that specification of the causal dynamics was seriously amiss, and that the subsequent results, no matter how internally eloquent, did not actually explain what happened. It "explained" what the perpetrators in retrospect said happened, but the historical evidence suggests that this was not the process that needed explaining. Milgram's work is the single best-known contribution within the experimental tradition in social psychology. However, it is an open question as to how much light it shed on what it purported to explain. In retrospect, Milgram's operationalization tended to reify the excuses that the Nazis gave at their trials (i.e., "just following orders"), making the murderers the victims of the Holocaust. Here, I am referring to the work of Daniel Goldhagen (1997) and Christopher Browning (1998). They reject Milgram's supposition that fear of authority was a prime factor in soldier compliance in the extermination of European Jews.[3] The evidence suggests that the Germans were willing executioners not fearful of authority, and that this is likewise true in recent ethnic conflicts in Africa and the former Yugoslavia (Brannigan 2013b).

There has also been a revival of debates about the foundations of intelligence with the publication of Herrnstein and Murray's *The Bell Curve* (1994). Where Rosenthal's *Pygmalion* (1968) argued that students' IQs responded to teacher expectations, the *Bell Curve* authors, Herrnstein and Murray, argued that the origins of variation were largely hereditary, and implied that such patterns might have a racial basis. One could not imagine approaches more diametrically opposed. Both studies purported to contribute constructively to discussions of public policy and both fueled debates that were, at times, incendiary. But what light has emerged from all the heat? *The Bell Curve* continues to attract condemnation for its apparent justification of racist policies (Yglesia 2018).

I think this inquiry is timely because it captures the chorus of voices from within the field who have registered a sense of both angst and ennui over its lack of progress. In my view, this chorus groans over the methodological focus that makes experimentation the lynchpin of scientific authority. My inquiry might suggest why this frustration is a natural outcome of the development of a psychology artificially confined within this methodological frame of reference. In this book, there is an assortment of problems that I am trying to sort out and for which I am going to offer some specific propositions or suggestions.

As Morton Hunt points out, the position of social psychology is puzzling. On the one hand, we must acknowledge the tremendous attraction of psychology to contemporary students. In North America and Europe, psychology is arguably the single most popular subject in the undergraduate curriculum, though its practical or career consequences are unclear. Even so, there is the recurrent fear among practitioners, as we saw in Chapter 1, that this subject is not actually going anywhere scientifically – this despite its orthodox devotion to the methods of Galileo and its tremendous attraction to its "end users," that is, the college students. This has bred periods of hand-wringing within the discipline that are never formally resolved. The hand-wringing is contradicted repeatedly by psychology's ability to address virtually any current social topic without developing cumulative theories with strong coherence and predictive validity. How can these various facts be reconciled?

Four propositions about the current state of affairs

My first proposition is that many of the classical experiments on which the credibility of the discipline is based are not experiments in the sense of the natural sciences. They are not tests in the strict sense, set up to compare outcomes on human subjects in experimental designs based on random assignment to different treatment groups. They are something else – they are demonstrations or dramatizations, not scientific tests. A demonstration has a lesson that is primarily pedagogical. It oftentimes contains a parable about everyday life. By contrast, a test is empirical and its logic is primarily "falsificationist," that is, designed to test the validity of a causal relationship against the possibility that no such relationship exists (Popper 1959). It seeks to build up a body of theory based on successive empirical observations, and its outcome is, in principle, aloof from moral or political preferences. Frequently, we confuse these two processes and treat demonstrations as tests – translating certain moral visions into facts. This problem is intensified because we emphasize the "objective" studies of things of greatest "subjective" relevance to us. This also happens in the natural sciences, but the process of replication there seems to put a check on the reproduction of empirically vacuous or ambiguous claims.

In social psychology, it is often possible for a claim to be honored even after its empirical foundations have been invalidated. The Hawthorne effect (as we shall see) and cognitive dissonance are good illustrations of this. This suggests that an idea may be more attractive than the evidence that supports it. How can that be so? Although there may be many different reasons, there is strong counsel in the field to ignore negative findings. Festinger advises:

> Negative results from a laboratory experiment can mean very little indeed. If we obtain positive results – that is, demonstrably significant differences among conditions – we can be relatively certain concerning our interpretation and conclusion from the experiment. If, however, no differences emerge, we can generally reach no definitive conclusion.
>
> (1954:142)

Why? The failure may simply arise from difficulties of achieving an effective operationalization – a point that, even if granted, seems to dispose of the null hypothesis *a priori*. Cold fusion was a great idea in physics but no credible source in the natural sciences would acknowledge that it has been established. In my view, social scientists following Festinger's counsel would not have been so skittish about the evidence.

If we allow that many of the classic studies are simply demonstrations, my second proposition is that many of the experiments borrow heavily from common-sense knowledge of social structure and simply iterate the obvious in an abstract methodological form. The experiment gives the sense of terrific scientific precision in the form of knowledge without actually discovering anything substantively new. However, it would be inaccurate to equate experimental knowledge with what is familiar and trivial. On the contrary, I hope to show that many of the great studies have a profound moral appeal that makes them qualitatively different from the experimental tests in the natural sciences. Why is this so?

This leads to my third proposition. The natural sciences are governed by a concern for the identification of things whose existence is more or less a question of fact: oxygen atoms have a weight of eight times that of hydrogen atoms, H_2O freezes and expands at lower temperatures that can be identified and measured. And natural science is concerned with relationships between variables – fluorocarbons either eat up the ozone layer or they do not. Pasteur's vaccinations either protect against smallpox or they do not. Genetically altered foods either make test animals sick or they do not. We can describe a methodology to determine whether these conclusions are valid and how much confidence can be attached to the evidence. This permits an assessment of the merits of the claim, particularly in the long run. However tricky it may prove to establish a claim in the short term (Latour and Woolgar 1979), it would be irrelevant if it could not be measured empirically and defended conceptually. By contrast, the experiments in social psychology are often motivated to demonstrate a condition that cannot be resolved by reference to facts and whose conceptual appeal is more subjective than objective. Many classic studies raise fundamental questions about human nature that are more the province of philosophy or divinity than empirical science.

My fourth proposition is that professional anxiety rises when these conditions appear. On the one hand, the discipline does not always follow the methodology of experimentation from which its call to legitimacy is based. And on the other, the sort of insight that is advanced as empirical knowledge is often not the sort of thing that accumulates like a body of empirical facts. Advances might better be described as contributions to philosophical anthropology or psychological ontology – disciplines intimately engaged in the experience of everyday life, but disciplines in which social judgments about historical experience are far more relevant than warehouses of transient facts or general theories.

Self-doubts in the discipline

Few academic disciplines appear to be as subject to self-doubts about their scientific achievements, prospects, and credentials as psychology.

> Anyone familiar with the broad field of psychology knows that it is in theoretical disarray. The different branches … proceed in relative isolation from one another, at most occasionally borrowing like a cup of sugar a concept here and a method there from a neighbor. Within each branch, psychologists also fail to reach consensus.
>
> (Buss 1994a:l)

This sweeping judgment was David Buss's opening salvo in a symposium devoted to evolutionary psychology, which he identified as a possible science of first principles in an otherwise fragmented discipline. As we shall see, Buss's recent misgivings about the state of psychology were not unprecedented. They are remarkably reminiscent of the earlier thoughts of Kenneth Smoke, who reviewed the contemporary state of social psychology in America. He noted that the textbooks varied enormously in what topics they covered, with the result that a reader conversant in one book could be "painfully ignorant" of the content of all the others:

> It might truthfully be asserted that [social psychology] is largely an amorphous mass, that in so far as it is able to formulate any generalizations, they are to be regarded as hypotheses rather than laws; that the worker in this field, unlike the physical scientist, is never quite sure whether he is studying stones or stars; … [and] that there is much metaphysical speculation in this field which is not ordinarily recognized as such.
>
> (Smoke 1935:541)

These two observations are not unprecedented reflections sitting like monstrous gargoyles guarding the beginning and the end of more than a half-century of experimental psychology. The misgivings about the scientific credentials of the field, especially in the case of social psychology, have been a recurrent subtext in methodological writings throughout the history of "scientific psychology" (Koch 1992a:7ff, Koch 1992b). Indeed, the "crisis in confidence" (Elms 1975) has propagated what many refer to as the "crisis literature." While many of the contributors to this literature have become outcasts from mainstream circles, other voices are at the core of the discipline: for example, Leonard Berkowitz, who stated: "Social psychology is now in a 'crisis stage' in the sense that Kuhn [1970] used this term" (cited in Elms 1975:967). Elms suggested that

> many social psychologists appear to have lost not only their enthusiasm but also their sense of direction and their faith in the discipline's future. Whether they are experiencing an identity crisis, a paradigmatic crisis, or a crisis in confidence, most seem agreed that a crisis is at hand.
>
> (Elms 1975:967)

Writing in 1980, Festinger commented about the field he left in 1966 in a piece entitled, "Looking Backward." He recalled that "much of the field seemed to me to be fragmented. Unfruitful disagreements and controversies arose all too often. New work that appeared could be quite ignored by others" (1980:247–8).

In the 1950s, when sociologists following Bales began to experiment on small group dynamics, the practice of operationalization that was integral to experimentation was criticized by Pitirim Sorokin in his classic discussion of "the illusion" of operationalism (1954). Sorokin decried what he called the "conversion" of laboratory psychologists to an "orgy of operationalism" in an attempt to mimic the success of the hard sciences. "The operationalists firmly believe in the infallibility of operational incantations [yet their] operational manipulations often resemble the 'scientific methods' of 'the scientists' in *Gulliver's Travels"* (Sorokin 1954:33). Sorokin's target was the "sham operationalism" that mimicked natural science methods with abstract jargon that overshadowed questions of context and meaning. For the most part, Sorokin was a voice crying in the wilderness as experimentation increasingly became the single leading methodology of scientific psychology and small groups sociology.

Irwin Silverman (1971, 1977) has chronicled a long list of similar misgivings about the logic of experimentation under the title, "Why Social Psychology Fails." He refers to Brewster Smith's 1972 review of the series *Advances in Experimental Social Psychology*: "Social psychology still trades more on promise than performance … We must conclude that the predominant experimental tradition in the field has contributed rather little for serious export in enlarging and refining our views of social man" (Silverman 1977:353). The following year, Smith wrote: "Our best scientists are floundering in search for a viable paradigm. It is hard to tell the blind alleys from the salients of advance" (1973:464). Rosenwald writes similarly: "Theoretical progress, as envisioned within the discipline of social psychology, is slow to arrive … even our laboratory-derived knowledge exhibits little of the cumulative character we associate with the scientific method" (1986:303).

Daniel Katz, in his closing editorial in *Journal of Personality and Social Psychology*, wrote that "the concern with technology and the marginal interest in theory are related to what seems to be the most critical problem we face today in social psychology – the continuing and growing fragmentation of the discipline" (Katz 1967:341). Similarly, Moscovici wrote:

> The fact is that social psychology cannot be described as a discipline with a unitary field of interest, a systematic framework of criteria and requirements, a coherent body of knowledge, or even a set of common perspectives shared by those who practice it. … From time to time the interests of the researcher are mobilized by themes or areas which appear new and important at the moment; but sooner or later these prove to be sterile or exhausted and they are abandoned.
>
> (Moscovici 1972:32)

Kenneth Ring (1967) wrote: "Experimental social psychology … is in a state of profound intellectual disarray" (cited in Silverman 1977:354).

Even the most enthusiastic proponents of experimental psychology sensed that the discipline was in trouble. Raymond Cattell opened the authoritative *Handbook of Multivariate Experimental Psychology* with the following observation: "Psychology is young as a systematic study and still younger as a true science. For many reasons its growing pains have been unusually severe and its progress fitful" (1988:3). Unusual growing pains? Fitful progress? Allusions to the novelty of the discipline are often cited as the reason for its disappointing progress. Even Smoke permitted that, despite its lack of focus, social psychology was "on its way," as though time would put the doubts to rest. However, this has not convinced R. M. Cooper:

> In the past any criticisms that have been voiced against psychology have most often been shrugged off as a reflection of its infancy. After 100 years or so this answer begins to wear thin, particularly given the exponential growth in activity of the past 20 years. One begins to suspect that psychology's failure is to be attributed not to immaturity but to retardation.
>
> (1982:265)

Cooper goes on to conclude that "my stance is that psychology is generally a failure. Every year psychologists turn out thousands of books and articles. I find it difficult, however, to see much in the way of fruits from these labors" (1982:265). Cooper, like Schachter (1980), believed that psychology faltered because psychologists failed to appreciate how biology controlled much of the behavior claimed by psychology.

Another gloomy confession was turned in by K. G. Ferguson in a piece entitled "Forty Years of Useless Research?"

> After almost 40 years as a student of clinical and abnormal psychology … I don't really know many more facts in the area than I did in the beginning. Watching the facts accumulate in one's field is worse than watching the hour hand on a large clock; you are tempted to wonder often if the clock has stopped.
>
> (1983:153)

Silverman writes with similar misgivings: "After more than three decades of progressive expansion, the social psychology establishment finds itself bereft of substance and direction" (1977:353). With respect to the textbooks, he draws conclusions reminiscent of Smoke: "Not only is there great diversity of content, but they are devoid of common definition of the field" (1977:354).

George A. Miller characterized the intellectual plurality in psychology as an "intellectual zoo" (1992:41). He wrote that

> no standard method or technique integrates the field. Nor does there seem to be any fundamental scientific principle comparable to Newton's laws of

> motion or Darwin's theory of evolution. There is not even any universally accepted criterion for explanation. What is the binding force?
>
> (1992:42)

Faith! "When reason fails, one resorts to faith … I believe the common denominator is a faith that somehow, someone, someday will create a science of immediate experience" (1992:42.) Note that these conclusions are not drawn by outsiders. They come from psychologists talking about their own discipline.

It is difficult to determine how representative such utterances are. Systematic probing of psychologists that explored their perceptions of crisis would be rife with social approval bias, and might reflect their sense of individual success as opposed to collective progress. The statements quoted here were not solicited by any investigator, and their credibility is enhanced by the fact that they were given at great risk of condemnation by colleagues. Many disciplines, particularly in the social sciences, experience misgivings about their scientific progress. In sociology, there are recurrent debates over positivism and the relative merits of quantitative versus qualitative approaches to methodology (Popper [1961] 1976; Cicourel 1964). As for economics, while outsiders may question the accuracy and predictive validity of macroeconomic models, this skepticism is not a view shared widely by insiders, with few exceptions (Learner 1990). The dismal science keeps economists quite buoyant about their intellectual credentials. But, in neither sociology nor economics do disciplinary limitations give rise to the sorts of recurrent frustration and soul searching witnessed in psychology. Psychology's plight would appear to be peculiar, and, in my view, its angst is most acute in the field of social psychology. Why should this be the case?

Why psychology?

Silverman captures a broad band of opinion when he argues that the adoption of the experimental method at the inception of the field was a Trojan horse destined to undermine the field's potential. "Social psychology became an institution solely on the basis of the vision that complex social phenomena could be fruitfully studied by experimental laboratory methods" (1977:355). The pitch was made effectively by Wundt in his laboratories at the University of Leipzig beginning in the 1870s, to which he attracted many American students who, in turn, imported his approach to North America, over the vocal objections of William James, the doyen of American psychology, and contrary to the influential work of G. H. Mead (1934). Social psychology developed under the wing of existing departments and, in the absence of any theoretical cohesion that would have dictated specific methodologies, it adopted, by default, the method that had supported Wundt's studies of perception and Pavlov's studies of physiological psychology, studies that went some way toward establishing the discipline's intellectual credibility. In a similar vein, Koch observes regarding the whole field that "psychology was unique in the extent to which its institutionalization preceded its content" (1969:64). Experimentation subsequently became the dominant medium of exploration, and this particular methodological focus inadvertently hindered the progress of its practitioners. Rosenwald, speaking of experimentation, put it in darker terms: “To cut through the Gordian knot, we have set our hopes and tightened our grip on one of the dullest blades available” (1986:328).

Experiments as short-term, low-impact designs

Silverman concluded that, after decades of experimental orthodoxy, psychologists were beginning to realize that “complex social phenomena cannot be fruitfully studied by experimental laboratory methods” (1977:353). Social psychological experiments are typically short-term, emotionally innocuous, low-impact designs calculated to have very little lasting effect on the subjects. However, many of the problems that interest psychologists, such as the causes of human violence and aggression, are not amenable to direct study in the laboratory. The social psychologist is forced to examine short-term analogues whose fleeting effects are measured immediately. Silverman concludes that this means that psychologists' generalizations are “never beyond the realm of speculation” in regard to their relevance to everyday life. “The conclusion which I draw is that experimental social psychology can never be serious” (1977: 356). There are exceptions. Some cases have resulted in high-impact work that has potentially traumatic and/or long-lasting influence on the subjects. Milgram's work is reported to have produced trauma in numerous subjects. Nicholson (2011) called it “Torture at Yale.” Zimbardo's prison study resulted in worrisome interpersonal aggression among those role-playing as guards. But these studies raise deep ethical questions and tend to be “one shot” affairs, exposing the researcher to professional criticism for risking human subjects and/or exposing them to harm on the one hand, while precluding the prospects of replication on the other.

Gadlin and Ingle similarly argued that the laboratory experiment was not always an appropriate format for the things that interested researchers. “Psychologists have begun to wonder about the external validity of the results of laboratory experimentation … Rather than selecting for research those phenomena suited to our methods, we ought to shape and develop our methods to fit phenomena” (1975:1003,1007).

Deception and ethics

Related to experimentation are two further issues that have worried social psychologists. The first concerns deception, the second concerns ethics. Because psychologists study actors who have a common stock of knowledge about society, the subjects are not naïve in the sense that animals in medical experiments are naïve about medicine, or electrons in science laboratories are oblivious to the laws of physics. As Orne (1962) has pointed out in his classic discussion of “demand characteristics,” subjects in experiments are role-playing, are seeking information from the environment, and fashioning their conduct in response to the expectations and information communicated to them, sometimes explicitly and sometimes inadvertently, by the setting.[4] This has raised questions about bias arising from "experimenter effects" (Rosenthal 1966). One remedy has been to make the subjects naïve by deliberately misleading them about the objective of the study. Social psychologists often devise elaborate cover stories to camouflage the object of their inquiry to ensure a "natural" response from subjects who, had they been informed directly about the point of the study, might have thought and acted differently. For example, Latané and Darley (1968) invited subjects to discuss problems of adjustment in large urban universities and, while they were filling in questionnaires, exposed them to what appeared to be an intense "accidental" discharge of "smoke" (titanium dioxide) from an air vent which led after several minutes to visual impairment and coughing. This was undertaken to study the bystander effect. Milgram told his subjects he was trying to determine whether punishment helped people learn, when in fact he was studying the role of authority figures in the mediation of aggression.

There are two concerns about deception. The first purely pragmatic concern is whether it succeeds. There has been some doubt as to how successful such deceptions actually are in keeping the subjects naïve (Strieker 1967). The issue is exacerbated due to debriefing. Since ethical considerations require the removal of the misapprehension or deception during the debriefing, there is concern that the cover story actually remains intact for future subjects after the initial subjects are debriefed. Although subjects typically are encouraged to keep their experiences to themselves and not to tell other potential subjects about the experiment, some evidence suggests that they do not comply, effectively undermining the cover for potential future subjects. In their replication of Schachter's anxiety and affiliation study, Wuebben, Straits, and Schulman (1974) found that a plurality of their subjects did not, in fact, respect the experimenter's counsel and conveyed enough about the study to potential new subjects to undermine the design. In addition, in Leon Levy's study, where the researcher's hypothesis was *explicitly* leaked to naïve subjects by those supposedly leaving the laboratory after the experiment, during the debriefing the majority of those in receipt of the insider information denied being told, despite the obvious impact of the communication on their performance (Levy 1974). The task involved verbal conditioning. Subjects were presented with cards on which verbs were printed and asked to make up a sentence using a pronoun. If the subjects chose "I" or "we," the experimenter reinforced the behavior by saying "good." Subjects in the experimental condition were met with a person who appeared to be leaving the setting having just completed the experiment. This was actually a confederate who said that the researcher was a PhD student worried about getting the right results, and that the subject had done "OK" once he figured out that he was supposed to say "I" or "we." The results showed a steep learning curve for reinforced behavior but the informed subjects started at a significantly higher level of compliance from the very start of the trials. During the debriefing, the vast majority of the subjects denied having been informed of the point of the investigation. These beneficent subjects feared spoiling the results and invalidating the research.

A second major concern with deception concerns ethics. If human subjects are deliberately misled about the nature of the research, in what sense can their decision to participate – which is voluntary – be informed? Consent presupposes what deception precludes. Though some researchers excuse the practice in terms of the benefits to society from the insights that result, others are uncomfortable with this instrumental logic, particularly as the returns to the investment in deception studies have been questioned (Baumrind 1964, 1985). Even if we allow a trade-off between short-term temporary deception and long-term scientific advancement, at what point is the profession prepared to present evidence of the long-term gain in knowledge and to justify deception? The claim to social benefits is gainsaid.

Others take a more absolutist view and argue that the systematic reliance on deception is categorically inconsistent with a professional treatment of human subjects irrespective of a general social good. However, without deception, the presupposition of subject naïveté is precarious. How does one strike a balance between full disclosure and informed consent on the one side, and normal (i.e., naïve) subject responsiveness on the other? Cook and Yamagishi (2008) argue that it should be avoided except where it is "absolutely necessary." Most behavioural economists think the practice is simply unethical. However one comes to a compromise, this raises an important substantive consideration in social psychology, which pertains to the use of experimentation with naïve subjects: the need for "naïvité" (and, hence, deception) arises because subjects are not by nature naïve about the motives and signals of others, including social scientists, and neither are they disinterested in them.

Common sense and scientific knowledge

What makes deception seem necessary in some quarters is the fact that people already have concepts and beliefs about the way society works independent of the leverage associated with an experimental method. Cattell attributed the "fitful progress" of the scientific study of human conduct at least in part to our pre-existing folk psychology. "Scientific writing has found it almost impossible to disentangle itself from semi-scientific, popular terminology, modes of reasoning, and 'theories' since 'psychology' is such an enormous daily preoccupation of all mankind" (1966:1). In a rare essay, Harold H. Kelley similarly noted the interplay between common-sense psychology and scientific psychology, and explored it at length. Common-sense psychology is "found under such rubrics as 'common sense,' 'naïve psychology,' 'ethnopsychology,' 'indigenous psychology,'and 'implicit theories'" (1992:1). If science and common sense are interchangeable, scientists have no claim to superior authority, and the utilization of esoteric methods of fact finding (i.e., experimentation) is superfluous. On the other hand, if the two fields of knowledge are more or less distinct, and if professional theories are more penetrating and reliable, resorting to specific methods of research is both desirable and necessary.

Part of the crisis in social psychology appears to have arisen from the observation that laypeople are already conversant with a lot of what passes for professional insight. John Houston's study of "lay knowledge of the principles of psychology" suggested that introductory psychology students and subjects contacted in a public park reported correctly on tests of cognitive processes at a rate far higher than one would have expected by chance. Houston: "A great many of psychology's basic principles are self evident" (1983:207). Manzi and Kelley reported a similar conclusion in a study of unequal dependence in pairs of couples: "The 'principles' we derive from the study of interpersonal relationships are already part of common knowledge" (Kelley 1992:3). However, Kelley was not prepared to conclude that there was no room for a scientific psychology. His essay was an attempt to clarify the mutual relationship between common sense and scientific psychology. What did he conclude?

Kelley observed that the interplay between common-sense psychology and scientific psychology was unavoidable because of the common linguistic and cultural immersion of psychologists prior to their elevation to the scientific frame of mind. Since scientific problems do not arise in a vacuum, common beliefs and terms "inevitably influence the concepts and theories we develop for our scientific purposes" (1992:4). Indeed, Kelley pointed out that there was often a strong correspondence between common-sense terms and scientific concepts, on the one hand, and between common-sense beliefs and scientific propositions on the other. Operationalizing common-sense terms like "commitment" or "closeness in a relationship" obviously borrows from everyday usage. However, an analyst may coin usages that are more precise or technical, and that organize observations in non-obvious, theoretically justified ways. The risk here is that this can eventuate in a level of jargon composed simply of pseudoscientific concepts. This was Sorokin's conclusion, alluded to earlier.

Prototype analysis

Obviously, Sorokin's concerns will temper any attempts to confuse common-sense understandings with scientific abstractions. Kelley identified a method for moving from common-sense terms to scientific concepts via what he termed "prototype analysis." Prototype analysis permits the theorist to poll ordinary language users to extract a family of interrelated concepts in both a "horizontal" and a "vertical" way, and, hence, to become more precise about their meanings. Horizontal variations tap different manifestations of a common idea: for example, *love* versus *caring* versus *protectiveness.* The vertical dimension reveals kinds of attraction (infatuation, liking, respect, etc.). Presumably, the analysis allows the researcher to reduce the ambiguities of a concept before entering it into a theoretical model. The transformation of common-sense terms into scientific concepts by prototypical analysis seems quite improbable. Indeed, the whole transformation of Wittgenstein's philosophy of language from the *Tractatus Logico-Philosophicus* (1922) to the *Philosophical Investigations* (1951) questions our capacity to put binding stipulations that can convert natural language concepts into definitive scientific terms. We may limit our scientific analysis to specific nuances of terms, although there is no guarantee that our readers and colleagues will similarly confine their readings to such nuances. To be fair, Kelley is skeptical about the project, noting that explorations of the "horizontal" dimension have outpaced the explorations of the "vertical" dimension (1992:11). Since higher order precision is based on vertical layers of meaning, this uneven development leaves open the question of the discipline's claim to a superior source of concept formation compared to common-sense theorizing. Kelley seems to sense this when he writes:

> I am expressing here some uneasiness about undue dependence on common thought for clues about how ψ-PSYCH should slice up its phenomena. There must surely be an important role for ψ-PSYCH analysis that enables our conceptual work to come partially under the guidance of logical and theoretical considerations and to avoid total dependence on common terms.
>
> (1992:12)[5]

If one were seeking a clear demarcation of the two realms and a lever to establish scientific concepts independently from common terms, although he surely believes these to be desirable, Kelley fails to identify them. The problem that Kelley tackles is the familiar charge that social psychology is the rediscovery of the obvious. "It reveals no new information, only what people already know" (Kelley 1992:13). Kelley counters that "what is obvious is not always obvious," particularly when viewed prospectively. He adds that common-sense beliefs are frequently "self-contradictory" – suggesting that ψ-PSYCH theories are not. The work of the psychologist, then. is to disentangle the conditions under which alternative outcomes arrive from similar premises, shifting attention away from large main effects to the more fastidious analysis of smaller interaction effects.

In a section titled "How to Make Science Interesting," Kelley suggests that a focus on the non-obvious should be a priority for psychologists since it generates interest in science. Kelley draws on the rhetorical analysis of Murray Davis (1971), who argues that "all interesting theories" dispute the "taken for granted world of their audience." However, Davis's position is that a lot of humbug passes as knowledge because of the way it is presented, and the powerful rhetorical tools of presentation are independent of the validity of the claims they convey – not that scientific innovations attract our attention because they are interesting. Hence, pursuit of the non-obvious character of the obvious may make psychology "interesting" in Davis's sense, but this has nothing to say as to the scientific relevance and value of such work. But the relevance may depend on the level of abstraction.

> Common beliefs are most likely to be veridical when they concern the mesolevel of behavioral phenomena, the familiar, and those events of which the person has principally been an uninterested observer … Beliefs

> about that behavior and its occurrence under various conditions should thus be fairly veridical.
>
> (Kelley, 1992:17)

Kelley is skeptical about common-sense beliefs when they move to the microlevel and the macrolevel – but these are levels of analysis typically outside social psychology, suggesting that the familiar territory of social psychology is situated where common-sense rules of thumb are already typically reliable ways of understanding the world, dispensing with the requirements of the special leverage of a scientific approach, and the specific methodology of deception and experimentation.

Kelley notes that many common-sense terms come with cultural implications that imply behavioral relationships. The psychologist who discovers such relationships empirically is basically only discovering how language works in the construction of reality.

> This suggests that CS-PSYCH can become a foundation for ψ-PSYCH theory. The creative work lies here in analyzing CS-PSYCH and revealing its underlying framework. Once any such theory is completed, we should hardly be surprised that, taken separately and viewed from the CS-PSYCH perspective, most of the specific ψ-PSYCH propositions will appear to be truisms.
>
> (Kelley 1992:21)

If that were the case, the interplay between the common sense and the scientific apprehension of reality would not be particularly fruitful. In fact, many would find it fatal. Kelley's own conclusion is ambivalent: "It is impossible for us to avoid the effects of CS-PSYCH, but easy for us to be unaware of them" (Kelley 1992:21). The problems arising from the role of common sense in scientific concept formation "deserve more widespread attention than they presently receive … The inevitable effects of CS-PSYCH on ψ-PSYCH are neither all good nor all bad" (Kelley 1992:22). The paradox is that the scientific analysis never gets far beyond common sense, and the subject matter of the discipline is, in Cattell's words, the "daily preoccupation of all mankind" (1966:1).

A consequence of this observation is that much important work in social psychology will inevitably be the common preoccupations of "lay psychologists" concealed in operational clothing, and dealt with abstractly through analogues. This makes social psychological experimentation a qualitatively different strategy compared with experimentation in chemistry or biology. The social psychologist studies, for example, how social norms form by showing how perceptual illusions are interpreted, or how mass media promote physical violence by studying how aversion to violence is taught in the laboratory. We are asked to treat the one as if it captured the other, but these are "as-if" analogies, not direct tests of social norms or physical violence *in situ* – and, for ethical reasons, this is unavoidable The upshot is that the psychologist as a cultural actor can speak to pressing issues of the day and can draw inferences about them as though they were based on the special leverage conveyed by scientific method.

In this view, experimental social psychologists are dealing earnestly with the perplexities of existence at arm's length through the use of experiments, but they are operating schizophrenically, often unaware of how their pre-theoretical knowledge guides their investigations. On the formal level, they are conducting tests and observations in the tradition of Galileo's discovering science, searching for new laws. Despite the fact that they never make comparable headway in identifying the *mathesis universalis* for the human sciences, the exploration is therapeutic. Like an analyst with society as a patient, it allows them to confront what troubles people in everyday life. The scientific progress of the experiment is illusory as science, but, at a deeper level, it contains an important unacknowledged subtext without which the ostensive work of inquiry would hold no attraction. Morton's "no man's land" has its finger on the pulse of society. It is the medium through which the scientist confronts the pre-theoretic perplexity of life, not directly, but obliquely, through the drama of the experiment.

Notes

1 *Walden Two* was the name of B. F. Skinner's 1948 utopian novel based on the principles of operant conditioning. Skinner's utopia was labeled after Walden, the 1854 account of Henry David Thoreau's two-year solitary sojourn at Walden Pond, near Concord, Massachusetts. Thoreau's ideas about the ideal life emphasized the need for a close contact with nature and freedom of the individual from unjust state interference. *Walden Two* was Skinner's attempt to sketch how post-Second World War Western society could optimize the "contingencies of reinforcement" to maximize human self-expression in an egalitarian, non-coercive society.

2 The term "existential determination" of science was coined by Karl Mannheim (1954) in *Ideology and Utopia* to refer to the cultural and other non-scientific influences on the content of scientific ideas. Rosenwald (1986) uses the term "extrascientific incentives" to identify the same process. Others refer to "pre-theoretical" knowledge to suggest how common sense influences "theoretical" or scientific knowledge.

3 There is some difference in approach here. Goldhagen (1997) argues that Milgram's experimental study was undertaken in a complete empirical vacuum in the sense that it presupposed that German soldiers were intimidated into complying with murderous orders. The examination of the war records of Police Battalion 101 in Poland suggests otherwise. No one was forced to murder civilians. The policemen complied with orders, were permitted to avoid executions, and in many cases volunteered for "Jew hunts" – the extermination of Jews who had run away from the ghettos to eke out survival in the forests. Milgram's portrayal misscripted the actual situation. People complied because it was ordered, and they acted differently from how they would have acted had the decisions been up to them – something probably true for the entire war effort. Browning (1998) is more sympathetic to Milgram, but not to the influence of authority figures. In fact, he points out that the authority figures in Police Battalion 101 were quite effete. Men complied for other reasons – peer pressure, careerism, loyalty to the unit, the perceived legality of the orders, and the like. Fenigstein (2015) provides a more contemporary assessment.

4 Martin Orne (1962) reported that the information-seeking characteristic of students recruited as volunteers in experiments employing deception was often responsible for

the main effects reported by the researcher. In his search for a baseline of normal compliant behavior needed for an understanding of hypnotic compliance, Orne found that subjects undertook the most inane challenges with gusto. He asked subjects to add up blocks of random numbers on numerous sheets of paper – and then to tear them up into pieces, and to repeat the whole process *ad nauseam*. He terminated the exercise after many hours since the students showed no sign of giving up. During the debriefing they communicated that they had figured out what was “actually” sought from them: perseverance! And once they got that into their heads it didn’t matter how apparently boring the job was since it actually amounted to a test of their character, and on that count, they were not going to let the experimenter find them deficient!

Orne generalized this conception of demand characteristics to the phenomenon of sensory deprivation effects. The conventional wisdom in the late 1950s was that sensory overload and sensory deprivation could result in a breaking down of ego coherence, that is, nervous breakdown. Sensory deprivation experiments suggested that subjects lost all sense of time, experienced hallucinations as well as intense bouts of anxiety, confusion, and fear. Because of this, subjects were asked to sign release forms waving their rights to sue researchers and their institutions for emotional damage. Orne replicated many of these diverse effects without actually exposing subjects to sensory deprivation. However, many of the trappings of the previous experiments – the conspicuous placement of a “panic button” for emergency rescue, the intimidating release form suggesting the possibility of hallucinations and emotional trauma, as well as the presence of a medical emergency station with appropriate medical personnel – conveyed the impression of risk so effectively that subjects experienced negative effects merely sitting alone in an office (Orne and Scheibe 1964).

[5] ψ-PSYCH refers to scientific psychology. CS-PSYCH is common-sense psychology.

3 Experiments as theater

The art of scientific demonstration in Sherif and Asch

Introduction: the classic influence studies

In the next two chapters, I elucidate the classic influence studies from the 1960s. What I hope to convey is that these remarkable investigations were undertaken with the idea that the experiment could be employed as a scientific device to illustrate a telling scientific fact. Strictly speaking, they were not tests to determine the validity of the fact, but a compelling means to dramatize it. The first cases arose from the work of Muzafer Sherif and Solomon Asch. These are followed by the later research of Stanley Milgram, Philip Zimbardo, and David Rosenhan.

The autokinetic effect and the liabilities of an illusion

In 1936, Muzafer Sherif published his classic study, *The Psychology of Social Norms*, in which he reported his elegant research on the autokinetic effect. The key experiments appeared initially in his 1935 article in the *Archives of Psychology*. The book was an attempt to publicize the work and spell out its implications in a much broader framework, and was followed by an extension of the paradigm to attitudes in the first volume of *Sociometry* in 1937. A norm is defined as “an authoritative standard” or model; “a principle of right action binding on members of a group, and serving to guide, control and regulate proper and acceptable behavior” (*Webster’s Dictionary* 1977). The basic design was quite simple. Sherif exposed subjects to a stationary pinpoint of light projected toward them in a darkened room. Subjects watched the light for a few seconds and were asked to estimate how much movement they saw, since it appeared to shift around in a ghostlike fashion. When asked to estimate how much the point of light had moved, subjects gave estimates similar to those reported verbally by other naïve subjects like themselves. These estimates converged. In contrast, estimates given by individuals privately without overhearing one another were disparate and independent, and tended to be stable between sessions and even across different days. In other designs, Sherif used confederates of some prestige to influence the estimates of naïve subjects. In addition, the experimenter in some cases directly suggested that the subjects were under- or overestimating the movements. The results showed that the naïve subjects tended to change their own range of estimates to bring them into line with the prestigious subject, and into line with the experimenter's cues. Sherif argued that the results were indicative of the process by which norms emerge naturally in society.

Sherif points out that the perception of movement is an optical illusion. The light appears to wander in the absence of a frame of reference. Consequently, the "group" effectively frames the individual's perception – as does the prestigious person or expert. Sherif acknowledges that this was well known to the Wurzburg psychologists from the previous century: "That aspect of the stimulus field is especially observed which the subject is set to observe" (1937:90). Presumably, external influence from others would set the field for the subjects. So, how do norms evolve? Sherif suggests that ego is influenced by others (in dyads, in groups, and by leadership figures) when the natural world is an ambiguous source of information. Individuals rely on one another to define reality when the environment fails to give clear clues. While no one would dispute this, it is doubtful that this is actually discovered empirically in the experiment. This would be more like a general supposition of empiricism as opposed to a hard-won fact established in the laboratory. We should also ask whether norms only arise under such conditions or whether they evolve in other ways. This would steer us toward a general theory of the evolution of norms – quite a tall order. In such a general theory, it would be natural to ask whether they also arise when the natural world is unambiguous. What does he mean to tell us by saying that this is how norms arise? This is not a historical study of specific norms. It is an investigation of norms at large. Yet, it is improbable that one could deduce that the behavior of several strangers watching a point of light, an optical illusion, could be indicative of the formation of norms in the sense defined earlier for a number of reasons.

There is no evidence that any reference group in a sociological sense existed in these experiments, no leader, no common history, no censure of misconduct, nor any of the usual things we ascribe to social groups. This does not dispute that there was *social* behavior. Certainly, strangers who spoke in one another's presence engaged in mutual turn-taking and reported similar estimates. However, the subjects' responses to Sherif's requests for estimates of movement would appear to be demand characteristics – ego hazards a guess since he or she *has been instructed to expect movement* (1936:95), and since it appears as though the others got away with similar utterances beforehand, and this, in a situation where it was really impossible for anyone to say for sure if there really was any movement, and, if there was movement, how much of it occurred. This interpretation – demand characteristics and external compliance – is an obvious consideration for anyone trying to replicate Sherif today. It would also be relevant to ask whether anyone thought the stimulus was an optical illusion and was not moving at all. Sherif recorded some of the impressions of subjects: "Darkness left no guide for distance. It was difficult to estimate the distance … There was no fixed point from which to judge the distance"

(1936:97). Significantly, Sherif acknowledges that "the effect takes place even when the person looking at the light *knows perfectly well that the light is not moving*" (1936:92, emphasis added). If that were the case, in what sense would this action be *normative*, since a verbal estimate measured repeatedly with great precision would correspond rather imperfectly to the subjective uncertainty recorded afterwards, and such exacting estimates could be given even though the subjects knew differently from what they saw. That raises another obvious point – whether the reported convergence was simply a conformity in reporting as opposed to an actual distortion in perception.

Another consideration is whether a series of strangers making the same ambiguous estimates – 100 times each in a round-robin fashion and repeating the process four times after short breaks for a total of 400 trials – constitutes a norm in any important sense of the term. There were no apparent consequences in terms of individual survival or error. Is an inconsequential assent to a number something that norms are made of? Norms are moral. We feel *compelled* to assent to the right answer. Optical illusions are perceptual. Does failure to comply lead to discredit or disorientation? What was the norm Sherif was really studying? Certainly, one relevant norm that escaped discussion seems to be that strangers accommodate the sometimes perplexing requests of psychologists even if they fail to make any immediate sense. Also, they seem to rely on one another's utterances where a failure to do so might make them stand out in a crowd. In fairness, Sherif acknowledges that the autokinetic effect "reduc[es] the process to a very simple form" (1936:99). This reduction of the phenomenon to a kind of decontextualized purity is at the heart of Sherif's use of experimental methodology. In my view, this methodological approach ironically undermines its specific empirical relevance while simultaneously giving it an air of complete generalizability. How could that be achieved?

Sherif writes that

> our whole point is that the autokinetic effect can be utilized to show a general psychological tendency and *not to reveal the concrete properties of norm-formation in actual life situations* ... Our aim is to show a fundamental psychological tendency related to norm-formation.
>
> (1937:93–94, emphasis added)

In other words, the way that people react to an ambiguous visual stimulus can be utilized analogically "to show" or help understand how social norms are acquired. But this must be done at the highest level of abstraction. Why? In the early chapters of *The Psychology of Social Norms*, Sherif lays the foundation for his approach by stressing that psychological inquiries often are biased by the "community-centrism" of investigators – the taken-for-granted social baggage that often clouds the perceptions of researchers by leading them to treat as normal quite idiosyncratic practices of their own cultures. The experiment is a method of putting distance and detachment between the experimenters and the objects of their own environments. The process by which individuals come to report perceptual displacement – something that is completely illusory and ostensibly not a question of moral preference – stands in place of the "concrete properties of norm-formation in actual life situations," which the researchers cannot tackle directly because of their own ethical moorings.

Sometimes, Sherif appears to presuppose that the processes of perceptual convergence and moral conformity occur in the same way, although the experiment is limited to the perceptual evidence. At other times, he seems to view them as quite distinct. The autokinetic effect is used as a stage to demonstrate, or dramatize, the larger and more important foundations of social norms that do not lend themselves to such easy exposition because of community-centrism. When he "venture[s] to generalize" from the basic lesson obtained from the experiment to social reality, he contends that "the psychological basis of the established social norms, such as stereotypes, fashions, conventions, customs and values, is the formation of common frames of reference as a product of the contact of individuals" (1936:106). The leap from perceptual convergence (which might be visual agreement, verbal compliance, or some mix thereof) to a wholesale range of normative structures appears like a sweeping revelation in the text, but is a halting *non sequitur* in logic.

For practical purposes, the conclusion that normative structures arise from "common frames of reference" would not shock many social scientists, but whether it would satisfy them theoretically is another matter. Also, few would draw such conclusions from the empirical evidence of the autokinetic effect. There are several specific problems. First, surely we can ask whether it is logical to argue from an optical illusion, an insecurity in visual perception, to social stereotypes, fashions, and customs. Are these not quite different things? Seeing something, agreeing that it has certain material attributes that can be described in common versus determining that it is socially desirable ("normative") are quite different kinds of judgments. Sherif enjoins us to conflate them.

Second, if we accept that "community-centrism" threatens the neutrality of scientific accounts, and agree that we need a general template for normative behavior (achieved via experimentation) in order to capture the common foundations of stereotypes, fashions, and customs, in what sense can the "common frame of reference" explain how norms arise, since its existence is already evidence of a normative foundation? In other words, it is tautological to explain the appearance of collective norms by virtue of a prior "common frame of reference," since this amounts to the same thing. The social contact between individuals that results in the common frame of reference constitutes, but does not explain, the rise of the normative order.

Third, despite the fact that experimentation, at least formally, is a deductive method in which the experimenter makes predictions about various outcomes based on differences in treatment, Sherif explicitly advocates a *post hoc* form of reasoning that is based on induction. For example: "If the principles established on the basis of laboratory experiments can be profitably extended to the explanation of the everyday operation of norms, then our principles are valid" (1936:68). And again:

> The test for such an approach lies in the applicability of the principle reached to the description and explanation of norms found in everyday life … Whether or not this is just one more psychological abstraction or laboratory artefact … can be decided after it has met facts in the fresh and wholesome air of actualities.
>
> (1936:88)

It could follow from this that if a researcher can discover an extrapolation to everyday life, that is what the experiment was essentially about in the first place – a position characterized negatively as *post hoc* reasoning in methodology but applauded as *serendipity* in theory construction. What makes Sherif's claim less of the latter and more of the former is his suggestion that his experiment was simply "an extension" (1936:89) of well-known prior observations in perception, specifically F. H. Allport's earlier work on group mediation of individual perception (1924:260–85). Also, Sherif reviews all the major gestalt psychologists on the ground–figure relationship (Külpe, Köhler, Henri, Wertheimer, and Koffka, among others). One could conclude that its relevance to the all-embracing conception of norms was arrived at in advance. This experiment was a demonstration, or allegory, designed to explain the general processes of interpersonal influence in everyday life, even if the explanation was more allegory than proof. As Sherif admitted, the autokinetic effect was only "the rudiments of the formation of a norm by a group … We have used laboratory material of a sort which is not found commonly in actual social life, but which, nevertheless, demonstrated the psychological processes in such cases" (1935:17, 47).

What was the existential problem for Sherif?

It is difficult to recover exactly what initiated Sherif's inquiries in the mid-1930s. He reports a concern for the dramatic changes in social life associated with the 1930s in America, the rise of totalitarian governments in Europe, widespread hunger and starvation, oppression of the powerless, and the mobilization of mobs through political sloganeering. He suggested that "the study of such unstable situations of oppression, hunger, and insecurity and their psychological consequences demand careful attention from social psychologists … especially in our time of transition" (1936:193). Again,

> When social life becomes difficult … the equilibrium of life ceases to be stable and the air is pregnant with possibilities … Such a delicate, unstable situation is the fertile soil for the rise of doubts concerning the existing norms, and a challenge to their authority.
>
> (1936:85)

Sherif was preoccupied with the important tensions in Europe and America that arose during the Great Depression, as were many of his generation. But, rather than tackle specific questions, such as the popular appeal of the Fascists in Italy and the National Socialists in Germany, he began by thinking about normative behavior in general, and dealing with the breaches in normative behavior in the 1930s at arm's length, as though he were examining a geometry of social relations in pure form, with idealized representations of people in general falling prey to unidentified sloganeering and irrational sentiments. Fear of bias from community-centrism made Sherif abandon the specifics of social reality, and the peculiarities of historical situations, in favor of treating everything as an expression of an underlying condition. The former would invite anthropological description while the latter could be analysed in terms of abstractions.

To study the social circumstances created when the equilibrium of social life ceased to be stable, Sherif turned to normatively neutral conditions easily operationalized in the laboratory: the autokinetic effect. To capture the origins of totalitarian norms, Sherif contrives a setting where subjects are scripted into roles that dramatize what the society has experienced at large. He suggests that subjects in a darkened room watching a phantom light actually perceive the movement that they report. And in the responses of these subjects to this illusion, Sherif himself sees the complexity of the society compressed to its essentials. He speculates on the evolution of normative behavior from watching people trying to figure out if a stationary light is perceived as moving a discernible distance as his subjects convey their impressions. Plato's allegory of the cave returns as the autokinetic effect. Berkowitz and Donnerstein (1982:249) have argued that an experimental setting does not have to have surface realism or demographic representativeness to be valid, a point that might recommend the value of Sherif's approach. However, as Baumrind notes, manipulations within specific experiments are so consequential for outcomes that "results do not survive even minor changes in the experimental conditions. … When the task, variables, and setting can have no real-world counterparts, the processes dissected in the laboratory also cannot operate in the real world" (1985:171).

Having registered some skepticism about Sherif's work from a purely methodological perspective, I also need to say that the story does not end there. As I hypothesized in the preceding chapter, psychologists often have a prophetic vision concealed in their science. Sherif does not disappoint us on this count. *The Psychology of Social Norms* is not a specialized treatment of group influence on visual perception. It is a brilliant and, at times, radical treatise on the very nature of social interaction, values, identity, and social change. Like other classical statements in the social sciences, it often blurs the line between ontology – the limitations and tragedies of human experience – and empirical inquiry. Among Sherif's suggestions, we find a call to end friction arising from class conflict – "the classes themselves must be eliminated" – and a call for the removal of "the belief in the divine origin of individual species" (1936:201) – both of which are "survivals" that create palpable harm to individuals. Viewed in this way, the autokinetic effect is what he calls a "prototype" of norms in this more expansive conception and becomes a vehicle for reflecting on the larger issues of human nature suggested by the social mediation of all our experiences, including sensory perception, by culture. Gardner Murphy suggests, in the 1965 reprint, that "the laboratory investigation presented to our faculty here is embedded in a matrix of social science considerations, nearly to the point of being completely lost … The details of the laboratory test have now become incidental" (Sherif [1936] 1965:x) – a point with which I agree. The empirical particulars were insinuated into larger considerations – both in the original endeavor reflecting Sherif's global interests in norms and attitudes, and in the subsequent focus on his methodology by proponents of experimentation.

Sherif's work became a classic study but the whole *post hoc* nature of his reasoning is never openly discussed, and the theatrical or dramatic structure of the experiment was similarly relegated to history. From this perspective, the suspicion that the subjects experienced no genuine shifts in perception resulting from group influence would be immaterial, since the process was already a matter of earlier scientific recognition. Some introspective accounts from the 1937 report suggest that certain individuals knew they were being influenced by others, although others apparently were influenced and either did not realize it or would not acknowledge it. After immigrating to the US in 1945, Sherif's work became far more concrete. His abstract laboratory experiments were succeeded by long-term field experiments.

Sherif, group conflict and the summer camp field studies: an archival exposé of Sherif's field experiments

After the autokinetic effect, Sherif's most famous work involved a summer camp experiment at Robbers Cave State Park in 1954 (Sherif, White, and Harvey 1955; Sherif 1956). This was actually his third field experiment. In 1949, he recruited twenty-four adolescent boys from New Haven, Connecticut. They were from under-privileged backgrounds and were invited for a "free" camp with the proviso that parents were barred from visiting the boys – to prevent "distraction". The parents and the boys were told that "new methods of camping were being tried out" (Perry 2018:29). The camp was held at Happy Valley in the Litchfield Hills in Connecticut. No one was informed that the conditions were being manipulated by psychologists to observe the group dynamics. After several days of normal camp activities where the boys socialized freely, they were divided into two groups – the Red Devils and the Bull Dogs. They were encouraged to bond in their new groups and the two groups were subsequently induced to compete for valuable trophies. This was done to illustrate Sherif's views of the origins of intense conflicts observed throughout the 20th century. These arose, not from the properties of individuals, but from competition between groups for scarce resources. In Sherif's view, it was possible to take well-adjusted individuals, to place them into arbitrary groups competing for scarce resources and, as a consequence, to produce intense hostility between them. The prizes were expensive, stainless steel jack-knives, and were highly coveted by the boys. After several days of competition, the winners were announced, and fighting broke out between the two groups. According to Sherif, on the last day at lunch, the two groups were "lined up on opposite sides of the mess hall calling names and finally throwing food, cups, table knives" at each other (quoted in Perry 2018:14).

Gina Perry discovered these details as a result of archival research of Sherif's papers that were donated to the Center for the History of Psychology at the University of Akron. For Sherif, the field experiments were more valid than those conducted in the laboratory because, unlike the autokinetic effect, they provoked genuine feelings and actions from experiences in real life. In the Happy Valley camp, he discovered how easily the subjects could be manipulated. Having created the conflict which his theory predicted, Sherif then began to think about how it might be possible to overcome conflict by creating a common interest where the competing groups were forced to put their differences aside to achieve a "superordinate goal." In other words, peace could be achieved by conditions that created common aspirations.

Sherif was awarded $38,000 from the Rockefeller Foundation to replicate the Happy Valley experiment with the addition of a new feature to produce harmony through the creation of a superordinate goal. He secured a camp facility in Middle Grove, New York in 1953, and recruited several assistants who would become quite famous – Marvin Sussman, O. J. Harvey, Jack White, and Herbert Kelman. The field experiment was designed to last three weeks and again the camp was cost-free with the stipulation that parents were barred from visiting. Through contacts with Protestant ministers, he chose twenty-four boys from the Schenectady area who were athletic with above average IQ, eleven years of age, from middle class, two-parent homes. This selection process standardized the potential role of age, social class, and religion, allowing him to maximize variations arising from group dynamics. In the Middle Grove summer camp, the boys were transported to the park together on a bus and initially bunked in a common mess hall. This would prove to be a mistake. Over the first few days, the boys made friends with each other, but then were arbitrarily divided into two groups and moved to separate tents, a process that proved very upsetting (Perry 2018:77–9). The roles of the research assistants were schizophrenic because they were basically observers and were tasked with taking copious notes about the boys' behaviors, listening to the boy's conversations in the tents before they fell asleep, observing their friendships, conflicts and emerging status differences. But, ostensibly, they were the adults and the camp counsellors. Everything was reported to Sherif after the boys retired. Some of the boys were suspicious about the microphones on the ceiling of the kitchen mess (Perry 2018:66, 130). The staff did little to discipline anyone. Sherif stipulated that status differences among the boys within the groups were to emerge naturally. In the eyes of some of the boys, the counsellors were not acting as adults. One of the several participants tracked down by Perry decades later noted that no one appeared to be in charge (2018:103).

The design of the experiment was three-fold. First, the segregation of the boys into distinct groups was designed to create a strong internal solidarity.

They became the Pythons and the Panthers. Secondly, the experimenters were supposed to create strong rivalry between groups through competition for scarce resources after the initial groups had coalesced with their own leaders and *espirit de corps*. And, finally, the groups were expected to make peace in the pursuit of a superordinate goal. However, many boys made friendships before the groups were segregated. Rather than coalescing as cohesive units, many group members became homesick. Sherif scheduled sport competitions designed to highlight animosities and fuel in-group cohesion. But often the groups supported members of the other teams. Sherif resorted to "frustration exercises" by having the men vandalize the property of one group who were expected to blame it on the other. Some of the counsellors secretly raided one group's tent and desecrated its flag to engender inter-group rivalry. This backfired when the suspects swore on a Bible that they had nothing to do with the event. "Ill-will between the two groups evaporated … any conflict had fizzled" (Perry 2018:124). The boys blamed the adults. "Kelman wrote that instead of directing their anger to their opponents, the Panthers turned on one another … the competition phase was supposed to bring each team together, but after each game there was recrimination and bickering" (Perry 2018:127). When Sherif wanted more pressure put on the boys to control their outcomes, Kelman noted that this was like the experimenter getting into the maze and pushing the rat (Perry 2018:132). Sherif switched out one of the counsellors because he was not doing enough to provoke animosity. The teams were then pitted against one another in a game of baseball for a prize of the stainless steel jack-knives where the stakes were winner-take-all. After Panthers defeated the Pythons, the winners insisted that the losers also had to be honored and they all had to shake hands. Good sportsmanship trumped feelings of dominance and submission. As Sherif observed how his theory was unravelling before his eyes, he and his assistants became deeply divided, pointing fingers at one another as to who was to blame. Sherif called off the experiment prematurely.

The following year, not having a final report for the Rockefeller Foundation, Sherif fielded a third camp experiment at Robbers Cave State Park in Oklahoma. This time, twenty-two boys recruited from Oklahoma City arrived separately and camped in isolation from one another. They bonded well, doing typical summer camp activities such as swimming, boating, and sports. They were then introduced into competitive tournament (consisting of baseball, tug of war, touch football, and tent pitching) which resulted in group hostilities – the Eagles versus the Rattlers – as expected. There was name calling, mutual raids on each other's tents in "commando style", burning of the opponent's flags, etc. But there was also a great deal of comradeship arising from sportsmanship. The superordinate goal was the subsequent creation of a water shortage because of the failure of the water supply, which required that the boys to work together as one team to remedy the crisis. This largely succeeded in producing the cooperation that Sherif expected. This led to his famous conclusion that "hostility gives way when groups pull together to achieve overriding goals which are real and compelling to all concerned" (Sherif 1956:58). That became the dominant narrative after the particulars of the second experiment were suppressed. But what Perry's account points out is that the adults played a critical role in orchestrating the conflicts. When one group raided another's territory, the adults acted indifferently. "It was the staff who kept the animosity going. The boys clearly looked to the men to police the misbehaviour of the rival group" (Perry 2018:185) – and they didn't. The Robbers Cave experiment was not a test of a theory "so much as a choreographed enactment with the boys as the unwilling actors in someone else's script" (2018:216). The cooperation over the superordinate goal was a return to normal levels of responsibility. The boys were "restoring rules that the men had broken" (2018:216). The radical non-intervention that resulted from a need to let social processes flow naturally sent a clear message to the boys. "The fact that [the men] did nothing to prevent or put a stop to the name-calling and the cursing, the food throwing, the vandalism, and the raids communicated their approval and encouragement" (2018:218). The researchers were like puppet masters pulling the strings behind the scenes to achieve the three-step process divined by Sherif's theory.

Perry's account raises important questions about this type of research. First of all, there was no informed consent for participants. The boys' parents and the church leaders used to recruit the subjects were told that the camps were recruiting participants to study leadership and character, to explore new methods of camping and other cover stories. But neither the parents, the boys nor those who helped recruit them were ever de-briefed after the experiments were over. When Perry contacted former participants decades later, they were dismayed to have been "used" in this fashion, and many had very mixed memories about their summer camp experiences. In addition, the camps exposed the participants to trauma that would have been less prevalent if the counsellors had acted less as impartial observers, and more as responsible adults. The experimenters also purposely engaged in activities in all three camps that were designed to offend the participants (sabotaging the boys' tents and fouling their clothing and other possessions) in order to provoke aggressive retaliation, and to nudge the boys to provide evidence that embellished Sherif's expectations. Unlike the autokinetic effect, this was a high impact dramatization based on deception by psychologists who did not seem overly preoccupied with the welfare of the young persons entrusted to their care for weeks at a time. The evidence suggests that the summer camps were used to illustrate Sherif's theories. Not to test them.

Asch and the resistance to social pressure

The work of Solomon Asch is the second classical contribution to social influence research in American psychology in this period. It appeared just after the Second World War. Where Sherif stressed how subjects were influenced by the group outlook, Asch was interested in the grounds of *resistance* to group pressure. He developed his ideas over the course of several publications (1951, 1952, 1955, 1956) that examined social pressure on individuals working in groups. He never referred to Sherif's field experiments. Where Sherif had studied social influence where the stimulus was inherently ambiguous and where individuals seemed to drift unconsciously into a consensus by exchanging opinions, Asch pointed to the social and individual conditions that compelled individuals to accept or reject opinions that they perceived to be *contrary to fact.* Sherif appeared to attribute the subject's knowledge to "the operation of suggestion and prestige" (Asch 1951:178). Asch stressed the predicament of the individual who can see differently from others but who experiences pressures to mimic them, and whose individual freedom is jeopardized as a result by a majority rule.

What was the existential problem for Asch?

As with Sherif, it is not clear what concrete issue initially motivated Asch's experimentation. On the one side, he appears to be making an intellectual response to Sherif. To be sure, he appears to have been involved in H. G. Sperling's MA thesis that replicated Sherif's work (in large part), and he devotes a significant portion of space in *Social Psychology* to a critical engagement with Sherif. Sperling's unpublished thesis was titled, "An Experimental Study of Some Psychological Factors in Judgment," and was presented at the New School for Social Research in 1946. It was reviewed by Asch at length (1952:487–90, 501) to challenge the validity of Sherif's paradigm. On the other hand, in all his publications, he stresses the context of propaganda and the manipulation of public opinion in the mass media. He proposes a basic study of interpersonal behaviors in order "to make fundamental advances in the understanding of the formation and reorganization of attitudes, of the functioning of public opinion, and of the operation of propaganda" (1951:177). He worries that the technical extensions of mass communications have created "the deliberate manipulation of opinion and the 'engineering of consent'" (1955:31). The final chapter of *Social Psychology* is devoted to the analysis of propaganda.

Social scientists were certainly aware of the enormously important role of propaganda, which had been so influential in mobilizing the German and Italian populations in the 1930s to support the war effort, and which, in the German case, promoted racial hatred resulting in genocide. Asch contested the Sherif paradigm that suggested that people tend to absorb their morality (i.e., norms and attitudes) from their social context. In Sperling's replication of the autokinetic effect study, subjects who were told that the stimulus was an optical illusion did not experience a drift of their estimates to a common range. And when they were exposed to a confederate whose estimates were wildly off, they did not feel compelled to absorb them in their own schemes because the other subjects appeared to be clearly mistaken in their views. For Asch, the individual's experience was a primary and independent source of information. His experimental designs focused on the dilemmas created when individuals had to confront vivid discontinuities between their views, and those of their neighbors. Sherif's subjects believed they were sharing the world known to them in common. Asch's subjects had to tackle the problem of defending what they knew to be true on the basis of their own senses, a situation that exposed them to potential ridicule and marginalization, or capitulating to the group and suffering a loss of self-respect and self-confidence.

In my view, Asch's experiment proceeds at two separate levels – the concrete manipulation of conditions and the development of the ontological condition of the pursuit of truth at personal expense. How could an individual stand up against misperception and false propaganda? Most readers will already be familiar with this telling study. What I would like to remind them of is that this work did not begin with a specific theory or hypothesis that the experiment was designed to test. It was another fishing expedition designed to explore Sherif's model based on an ambiguous stimulus with an alternative social pressure that was downright provocative. Asch's device for exploring this interest in the laboratory was to ask subjects to match the length of a stimulus line that was drawn on a cardboard sheet with one of three other lines represented on another sheet. Although the correct match appeared highly self-evident, the unsuspecting subject found himself sitting at the end of a round-robin of guesses from seven to nine others and at odds with them in a third of all the guesses. Unknown to the real subject, the confederates were instructed to choose incorrectly. Many subjects were completely floored by the situation and removed their glasses to "double-check" the stimulus board. In about one-third of all the critical trials, subjects mimicked the majority. Three-quarters of subjects were swayed at least once by the erroneous majority. A third of the subjects caved into pressure at least half the time. Nearly all were emotionally provoked by the inconsistency.

Of what social situation is this an operationalization? Propaganda? Public opinion? Advertising? It is hard to say. It is presented as a generalized investigation of social influence. The most intriguing findings are that subjects show tremendous variation in their responses, some acting independently throughout, and others caving into group pressure at every turn (Griggs 2015a). Asch devotes a considerable discussion to variations in how subjects reacted to the confrontation that the design produced.

Asch introduced several variations to the basic design to determine the effects that these had on the levels of influence: the presence and absence of an ally, effects of changes in the size of the majority group, and effects of variations in the degree of the group error. What was discovered? Under what conditions do people resist propaganda or other social influence? Asch discovered that errors made by real subjects following group pressure to err declined when *one other subject* chose correctly. Ergo, external pressure is resisted when ego has an ally! But the ally must be constant, for if the ally bails out midway through the trials or arrives late, ego's vigilance for truth declines accordingly. As for the size of the group, a maximum influence occurs with a majority of three. Larger groups exert no higher levels of conformity. And finally (contrary to Sperling), there was little evidence that subjects ignored majorities that reported large errors as opposed to moderate errors.

Just as Sherif tackles social influence allegorically, Asch's entire orientation appears to have little relevance to everyday life. I would again hazard an opinion that this experiment tells us nothing informative about the process of propaganda during the Second World War. It says nothing about genocide and the political use of scapegoats to misattribute the real misery of German society during the 1930s. It says nothing about national animosities, nor the state's legitimation of violence to deal with opponents. Like the Sherif experiment, it borrows from the pre-theoretic understanding of the phenomenon of propaganda in order to context the experimental task of line discrimination.

It is ironic that, though the experiments arise from pressing issues in the life world, protocol dictates that this social relevance be studiously misrepresented in the experiments themselves through a deceptive cover story to prevent the subjects from learning the point of the study – in this case, to study the effects of propaganda – and to prevent them from acting on this definition of the situation, presumably to consciously resist it. Where the consumers of the experiment orient to history for its relevance, the actors or subjects must operate in the dark so as to recapitulate history from the stance of naïveté – again demonstrating that the experimenter can release or bottle up the phenomenon as required. As in a box camera, things are turned upside down as our grasp of the world is used to explain and make sense of the experiment in the laboratory, and as the same life-world relevance is hidden from the subjects to ensure they do not invoke their own common stock of knowledge of the world to exhibit how propaganda ought to be dealt with.

Asch's moral agenda: resistance and conformity as ontological dilemmas

Students of experimental social psychology seem to ignore the fact that the particulars of some of the classic studies are empirically vacuous. Does anyone really believe that Asch discovered a critical number (i.e., three) that results in maximum social influence in group situations? To which settings could such a discovery be generalized? Or that the role of an ally or friend in opposing false knowledge is any more than what one would guess from common sense and no more or less reliable? Would anyone build an organization based on these specific findings? I believe that would be foolish. Indeed, from an empirical perspective the research is quite casual. There is no pretense that the subjects are representative since they are acquired through snowball contacts. There is no control for gender. The reports also differ significantly in their particulars. The 1951 chapter reports that there were eighteen trials with twelve critical tests involving a total of eighty-seven subjects. In, 1952, Asch reports fifty-six subjects involved in twelve trials of which seven are critical. And, in 1955, the number of subjects jumps to 123 in eighteen trials. What gives? One is reminded of Harold Garfinkel's experiments in *Studies in Ethnomethodology* (1967), in which he tells readers that his experiments are "aids to the sluggish imagination," that is, demonstrations, and not to be taken too literally, advice that seems equally appropriate here.

I think readers of Asch overlook such details because the description of the predicament of the subjects is so engaging, and the analysis of their situation makes a point that transcends the original study, although the point is more philosophical than empirical. Some subjects acted with courage and confidence in confronting their situations but most were deeply threatened and disturbed, oftentimes experiencing a "double-take" to confirm that their neighbors were so clearly wrong, sometimes laughing nervously and sometimes withdrawing. Most subjects erred to some extent during the critical trials. Among the yielders, only one subject said his perception of the lines changed after he heard the majority opinion, although, as Asch notes, "we cannot be fully certain of what took place" (1952:469), and certainly he does not think any of Sherif's subjects actually experienced distortions in perception. More likely was the situation where subjects caved in for "the fear of exposing themselves to ridicule" (1952:470). Even if they enjoyed a short-term relief from embarrassment, they subsequently experienced a feeling of personal defeat and were racked with feelings of self-doubt and helplessness.

Of what relevance are these issues? Asch frames the problem as an inevitable condition of the social order. Social order requires a degree of consensus for the operation of group life. People enter into social relationships with a certain amount of trust in the value of those relationships but independence is also necessary at both the individual level and the collective level. At the individual level,

> to be independent is to assert the authentic value of one's own experience; to yield is to deny the existence of one's senses, to permit oneself to become confused … to renounce a condition upon which one's capacity to function depends in an essential way.
>
> (1952:497)

At the social level, the act of independence is essential to prevent the spread of errors and confusion. "The meaning of consensus collapses when individuals act like mirrors that reflect each other" (1952:495). Asch says that he "cannot rigorously justify the relevance of the present observations" to the general social conditions that people face, but there is good reason to believe that this juxtaposition between independence and yielding is a central dilemma in social life.

> There are times when one must choose between stark alternatives that have very much to do with the question of independence. Germans who lived near concentration camps could not escape the choice of breaking with their social order or of forcibly suppressing a range of facts and refusing to bring them into relation with their daily experiences.
>
> (1952:496)

So, Asch does tackle propaganda – but at arm's length, by classifying people as independents or yielders. Yet this only idealizes social processes and does not throw much light on actual, historical experience.

Final point. Asch's work has some of the trappings of experimental manipulation of the conditions of influence (role of ally, group size, etc.) but generalizations from these would be trivial, if not reckless, especially as the evidence is inconsistent (i.e., of the magnitude of group error). The most interesting part of the study is inductive – the identification of ideal types of reactions among "yielders" and "independents." Now we ask, how does Asch explain this polarity in social life and which trait comes to dominate in an individual? One of the less well-known positions that Asch advanced was that such traits tended to be relatively stable across situations. Such traits were marks of "character" for Asch and were not readily amenable to investigation through experimentation. He rejected the idea that the differences in question were "constitutional" (i.e., innate) and suggested instead that "the present discussion converges on a difficult and intriguing problem: the relation between character and social action" (1952:499). He seems to imply that certain social conditions will better foster independent action and build community consensus by drawing on the mutual dependence of personal and social qualities. It is interesting that when Asch's 1951 paper was revised for inclusion in Proshansky and Seidenberg's (1965) popular edited collection, *Basic Studies in Social Psychology*, the final sentence on "the relatively enduring character differences" he identified was deleted by the editors. This forced greater attention on the experimental variations of the work, but highlighted findings with the least scientific relevance. The most important element that is stressed in all the reports of the line discrimination task is edited out, presumably because it did not lend itself easily to experimental investigation.

4 Scientific demonstration in Milgram, Zimbardo, and Rosenhan

More evidence from the archives

Introduction

In this chapter, we focus on three of the most provocative studies in classical social psychology: Stanley Milgram's obedience study, Philip Zimbardo's Stanford Prison Experiment, and David Rosenhan's study of psychiatric hospitals. The theme that unites these diverse investigations is the utilization of the experiment as a pedagogical device to demonstrate a perspective whose findings are a foregone conclusion.

The Holocaust and obedience to authority

Although Milgram's study was derived conceptually from the work of Solomon Asch (Sabini 1986), the trial of Adolph Eichmann sharpened the issues for him. Eichmann was the allegedly plodding Nazi bureaucrat who assisted in the mass murder of European Jewry by masterminding the concentration of the victims in Poland after the Nazi occupation of France and most of western Europe. Subsequently, the Nazis developed factories for the extermination of Jewish victims at Treblinka, Sorbibor, Auschwitz, and other death camps. Several million innocent people, men, women, and children, were murdered at these death camps by ordinary German administrators, policemen, soldiers, and camp guards. In Milgram's experiment, ordinary subjects were cast in the parts of executioners. In the "received view" of this work (Stam, Radtke, and Lubek 1998), Milgram took people from all walks of life and turned them into the experimental analogs of Eichmann, suggesting that the capacity for evil was fostered in virtuous individuals by monstrous bureaucrats. The existential problem could not have been more clear-cut. Indeed, all of Milgram's work has the bite of immediate relevance.

The study was advertised as an experiment designed to test the effects of punishment on human learning. Subjects ("teachers") were paid to teach the "learners" to memorize a long series of paired associations. The pretext for the study was to advance knowledge about the effectiveness of negative reinforcements on learning. Errors were to result in a shock, but the level of the shock escalated at every mistake in fifteen-point gradations from 15 volts right up to and beyond 450 volts. The experiment was run with individual teachers and learners, but the role assignment was rigged so that the real subject was always assigned the role of the teacher who administered shocks, while an affable middle-aged man, a confederate, acted as the learner. The teachers were given a sample shock to demonstrate the actual discomfort that resulted from their control of the shock machine. The machine was an impressive electrical appliance with switches, lights, and verbal designations describing the severity of the shock (mild, moderate, high, extremely high, XXX). The subjects were drawn from a wide range of occupations and professions, unlike the usual captive population of undergraduate students.

The single feature of the research that advanced the influence studies was the utilization of an authority figure who appeared to be the scientist directing the experiment. His job was to pressure the teachers to comply with demands to administer increasingly severe levels of shock (which Milgram equated with aggression), since the learning task was rigged so that the learner's performance attracted increasingly painful (but illusory) levels of punishment. The experiment produced tremendous anxiety in many of the subjects.

Stanley Milgram wrote:

> Many subjects showed signs of nervousness in the experimental situation, and especially upon administering the more powerful shocks. In a large number of cases the degree of tension reached extremes that are rarely seen in sociopsychological laboratory studies. Subjects were observed to sweat, tremble, stutter, bite their lips, groan and dig their fingernails into their flesh. These were characteristic rather than exceptional responses to the experiment …
>
> … One observer related: "I observed a mature and initially poised businessman enter the laboratory smiling and confident. Within 20 minutes he was reduced to a twitching, stuttering wreck, who was rapidly approaching a point of nervous collapse".
>
> (1963:375)

Like the previous studies in this tradition, there was no *a priori* identification of hypotheses, nor specific examination of alternative theories. More fishing. Ties to Sherif and Asch were absent. Milgram cast his work as though it were generated *de novo* without influence from the earlier research. Milgram approached many groups to determine what they thought would be the normal responses to his experimental manipulations, and, especially, what people would estimate the refusal rates would look like. Psychiatrists, college students, and middle-class adults predicted that 100% of the subjects would defy the authority figure and refuse to administer the lethal levels of shock.

In the *Blackwell Reader in Social Psychology*, Hewstone, Manstead, and Stroebe summarized the study: "There is no experimental design as such; no factors are manipulated. No statistics are reported on the data nor are they needed since no experimental variations were compared" (1997:54). This characterization is not entirely fair. Milgram studied a number of different conditions of aggression, the most famous of which was proximity. He argued that the closer the victim to the context of aggression, the lower the levels of compliance. He also tested the effects of group mediation of compliance. Indeed, he reports twenty-three different conditions of obedience, suggesting again that the research was inherently inductive. Milgram found that the majority of subjects in the baseline experiments *did* administer the maximum level of shock but that this declined the more proximal the victim was to the teacher. He concluded that compliance of individuals in bureaucratic condition results from the force of authority figures on their obedience. His experiment extracted this general human tendency from the reports of the Holocaust killers who reported that their role in genocide was a result of "just following orders." This has been the dominant view of the obedience studies over the last six decades.

Criticisms were raised both in terms of internal and external validity. As for internal validity, contrary to the received view, Orne and Holland (1968), Mixon (1971), and other critics argued that, in psychology experiments, subjects presume that "nothing can go wrong" and that bad things may not be as bad as they seem. Even though subjects are told that the shocking device delivered some 450 volts and are demonstrated through a sample that the volts are, well, electrifying, most presuppose that "this must be OK – no one can really get hurt." Universities cannot permit that to happen.

In the pre-tests of the study, Milgram reported that "in the absence of protests from the learner, every subject in the pilot study went blithely to the end of the board" (1974:22). Meaning what? Every subject in the pre-test administered the maximum shock level without pressure from anyone. No one stuttered, sweated, or shook with anxiety. It was only at this point that Milgram introduced the various feedback conditions – initially a knock on the wall to indicate that the learner receiving the shocks was actually experiencing discomfort. In the *Obedience* film, it is evident that when the fake learner exhibits pain by actually shrieking – along pre-recorded lines – the real subjects *initially* laugh out loud. They are *startled* that anyone is actually being hurt. In the later designs, when the subjects hear similar complaints from the learner testifying to the painfulness of the shocks, they also have in their presence the "authority/scientist figure" – the actor-experimenter who contradicts their perceptions that something is going wrong, and who reacts passively as people appear to be suffering nearby. The subject is drawn between what is heard – a suffering victim – and what is seen – a nonplussed authority figure subject to the same information but not alarmed by it. This causes enormous conflict for the subjects. They frequently sweat, stutter, and tremble. They are mortified by evidence that the learner is suffering. This is a rather different scenario from the Eichmann episode where the stench of death in the camps was unmistakable. Neither does this dispose of the Orne and Holland critique. People may have started with an assumption that nothing can go wrong only to have this contradicted by what they could hear from the learner, but not by what they could see from the authority/scientist. As Orne and Holland (1968:287) note:

> The most incongruent aspect of the experiment … is the behavior of the Experimenter … Incongruously, the Experimenter sits by while the victim suffers, demanding that the experiment continue despite the victim's demands to be released and the possibility that his health may be endangered. This behavior of the Experimenter, which Milgram interprets as the demands of legitimate authority, can with equal plausibility be interpreted as a significant cue to the true state of affairs – namely that no one is actually being hurt.

The credibility of the experiment is not furthered by the fact that the role of the teacher is actually superfluous in the experiment, since the teaching could obviously be carried out without volunteer teachers. In the same vein, it could not have escaped notice by all the subjects that the learning task was simply impossible, and the demands quite incredible. This was Mantel's observation (1971:110–11):

> Every experiment was basically preposterous … The entire experimental procedure from beginning to end could make no sense at all, even to the laymen. A person is strapped to a chair and immobilized and is explicitly told he is going to be exposed to extremely painful electric shocks. The task the student is to learn is evidently impossible. He can't learn it in such a short space of time… . No one could learn it … This experiment becomes more incredulous and senseless the further it is carried.

Mantel is often cited as someone who "replicated" the experiment but his own views about its ecological validity, that is, its relevance to everyday life, are often overlooked. In a similar vein, Baumrind noted that

> far from illuminating real life, as he claimed, Milgram in fact appeared to have constructed a set of conditions so internally inconsistent that they could not occur in real life. His application of his results to destructive obedience in military settings or Nazi Germany … is metaphoric rather than scientific.
>
> (1985:171)

Don Mixon suggests that every experimental manipulation that Milgram developed which introduced less ambiguous evidence that a subject was being hurt reduced the aggression of the teacher. When the learner's pain was signaled through pounding on the wall, compliance dropped from 100 to 65%. This was the single most significant variation tested. It is also another fact lost on the textbook writers. All the elaborate verbal feedback of learners' suffering that was used as the baseline treatment reduced the compliance by only a further 2.5% over the knock on the wall – meaning that only one less person in forty resisted going to the highest shock level. Even though the authority figure is central to the received view of the study, his inclusion was actually a later addition to the design. Milgram thought that the verbal designations on the shock levels written across the electrical device would impede obedience on its own. That it did not suggests that people did not expect suffering to come to citizen volunteers. The classic study only emerged when he introduced feedback of harm and equated compliance with a specific agent – the lab-coated scientific "boss." But surely this was illogical, since the "harm" occurred at highest levels without the expert authority. The introduction of the latter contributed not power over the subjects as much as ambiguity over the harm.

From this perspective, the study is an inductive exploration, not a deductive test of theory. When the victim's suffering was brought into the room and portrayed dramatically by an actor in the real subject's presence, although the authority figure's comportment suggested no harm, the aggression declined. And when the authority figure was totally removed from the laboratory, the pain feedback information reduced the shocks to extremely low levels. In other words, the more evident the painfulness of the procedure to the innocent teacher and the more the background expectation that nothing can go wrong was contradicted by experience, the lower the levels of compliance to the authority's demands. In a *post hoc* questionnaire completed by 658 former subjects, only 56% suggested that they fully believed the learner was receiving painful shocks. This involved less than half of the obedient subjects (48%) and most of the defiant subjects (62.5%). Over 40% were unclear as to what they perceived. So, it is not clear that the manipulation was nearly as successful as the "received view" suggests, and when subjects *did* perceive harm, they tended to be defiant. On this reading, the experiment should have been grounds for optimism about humankind. Milgram did not throw any more light on the subject matter than was already evident from history. The final frames of the *Obedience* film depicting the pulsating force field of the authority figure – crudely tying his work to Lewin's (1951) field theory – end with a warning more appropriate to vintage science fiction movies. "In comparison to the effects tested in our New Haven Labs, one can only wonder at the altogether more powerful influences wielded by governments and bureaucracies on individuals."

This would *already* have been self-evident to any student of the war. It was, for example, laid out in Shirer's brilliant report furnished so quickly after the end of the war based on his diplomatic and journalistic coverage of the events. Compliance in war crimes by whole police and army regiments was documented at the Nuremberg trials. As for Milgram's contributing anything of theoretical significance, the experiment was a theoretical cul-de-sac despite the massive public attention devoted to it. Milgram (1974) argued that persons who entered a bureaucracy slip into an "agentic state" which expunges their autonomy. The "agentic state" is as tautological today as it was when invented. Indeed, Milgram only speculated about the state years after the experiments were finished (Blass 1992:279). It is alarming to think that the study that attracted more attention than any other in its generation did not result in any novel, theoretical insight. The experiment's extra-scientific attraction was simply this: it allowed the psychologist to dramatize the story of humankind's capacity for ruthless violence in an experimental idiom. By replaying Eichmann in the laboratory, it did not substantially advance knowledge, and neither did it discover anything essential or new about the death camps. Furthermore, it obfuscated the deep anti-Semitism that fueled the destruction of European Jewry by the Nazis, and substituted generic obedience.

In contrast to the view suggested by Milgram, the recent revelations of historians Daniel Goldhagen (1997) and Christopher Browning (1998) suggest that ordinary Germans were overwhelmingly complicit and willing participants in the slaughter of the Jews. Members of the police battalions who carried out many of the initial mass shootings who asked to be relieved from the killing were reposted without recriminations. Furthermore, many executioners inflicted suffering and humiliation on their victims far beyond what was ordered by the state.

Hannah Arendt (1964) stresses points in the evidence that Milgram seems to miss. The first was that the policy of genocide was the rule of law in Germany during the Nazi period. In other words, like the Allied carpet-bombing of German and Japanese civilians, killing of non-combatants was based on the rule of law at the time, however repulsive it was in its consequences. Milgram's conceptualization seems to depict the Germans as unwilling executioners, contrary to the historical accounts of Goldhagen (1997) and Browning (1998). In transporting these issues to the laboratory, Milgram's design is based on the supposition that the teacher's aggression is not only illegitimate, but is seen to be illegitimate by the subjects (by implication suggesting that ordinary Germans did not participate in genocide except against their wills). But this conflates two rather different contexts. Subjects have learned from childhood that it is a fundamental breach of moral conduct to hurt another person. But during war, do not most people believe that it is morally appropriate to kill to ensure collective survival in self-defense? While people may not like it, failure to do otherwise could cause one's own death. Also, Milgram's stipulation about what people have learned from childhood seems oblivious to the realities of intergroup hatreds that systematically reduce altruism and escalate intergroup conflict.

There is further moral jury-rigging in Milgram's account, identified by Patten (1977a,b). If Milgram knew during the course of his experiment that subjects were being hurt (i.e., emotionally traumatized), why did he not terminate the experiments immediately? Answer: he thought science might benefit in the long run. However, in characterizing the conduct of his teachers as acting in a "shockingly immoral way," Milgram overlooks the fact that the subjects might be entitled to the same excuse since, during the cover story, they were encouraged to administer electric shocks to advance human knowledge about the effectiveness of punishment. If acting to advance science, would the subjects characterize their conduct as "immoral aggression" (bad) or "reinforcement" (good)? Milgram has it both ways. He describes the task to subjects as a legitimate exercise, then characterizes it as immoral – oblivious to the parallels with his own callousness toward the subjects. Abse suggests that if one wants to view the subjects as so many Eichmanns, then "the experimenter had to act the part, to some extent, of a Himmler" (1973:29). Even if we disagree, we must acknowledge the double standard.

Arendt's second major point was that, during the Nazi regime, the policy of "resettlement" and genocide would have been impossible without the cooperation of the Jewish leadership, something Eichmann identified as a "cornerstone" of Nazi efficiency. However, the role of the victims and the capitulation of leaders who betrayed them are simply omitted from the experiment. While recognizing that these considerations might present formidable design questions for the experimenter, a failure to tackle them meaningfully has the result that in exploring one of the darkest pages in Western history – and attracting our interest for this very reason – Milgram's experiment boils it down to Punch and Judy simplicity: bureaucracy made good people behave badly. Unfortunately, when we come to other cases of genocide, such as the mass shooting of Vietnamese villagers at My Lai, Milgram enjoins us to read it as just another case of the crime of authority. In comparing more recent atrocities with the Nazi massacres, Goldhagen (1997:14) offers an alternative view:

> Who doubts that the Argentine or Chilean murderers of people who opposed the recent authoritarian regimes thought that their victims deserved to die? Who doubts that the Hutus who slaughtered Tutsis in Rwanda, that one Lebanese militia that slaughtered the civilian supporters of another, that the Serbs who killed Croats or Bosnian Muslims, did so out of conviction of the justice in their action? Why do we not believe the same for the German perpetrators?

Milgram's moral vision: on human nature, fate, and violence

One of the great attractions of Milgram's work is the latent moral agenda that surfaces in the final pages of his 1974 book. Milgram suggests that the inability of individuals to resist the pressure from authority figures is inherent in our make-up as a species, and as such, represents a design flaw that could jeopardize our survival. The subjects in the obedience experiments acted with violence against an innocent person, but not out of anger or provocation. "Something far more dangerous is revealed: the capacity for man to abandon his humanity, indeed the inevitability that he does so, as he merges his unique personality into larger institutional structures" (1974:188). This "fatal flaw" gives our species "only a modest chance of survival." Human nature "cannot be counted on to insulate" people from "brutality and inhumane treatment at the hands of malevolent authority" (1974:189). A substantial number of people will follow genocidal orders "without limitations of conscience, so long as they perceive that the command comes from legitimate authority." Ignoring, for the moment, Milgram's conflation of *malevolent* and *legitimate* authority, this vision of individuals fated inevitably to absorption by institutions, unprotected by a transcendental conscience, and, by nature, prone to violence and mutual destruction is Promethean in its scope. The psychologist as prophet reads the Holocaust only as an instance of this more pervasive condition of humanity that stems from the very core of our being, and that bodes ill for the future of the species.

As with Asch and Sherif, the moral tone is miles from the evidence, and ignores the methodological limitations that exercised the critics. But the ethical appeal is undeniable. It recapitulates the story of genocide stripped of its historical particulars and depicts it as an expression of a side of human nature that cannot be redeemed by conscience. Surely there's a lesson there. Despite the official orthodoxy, experiments serve as platforms for the dramatization of ideas, not for the testing of hypotheses and the building of theories. And that seems unlikely to change, given the centrality of the experiment in the arsenal of social psychologists. But the moral tone also explains the enormous appeal of the field to students and the public, who get an "ethical fix" packaged as science, and who enter the moral high ground under the guise of scientific training.

Recent archival re-assessment of the Milgram conclusions

Australian writer and psychologist, Gina Perry, spent four years researching the Milgram archives at Yale University. Over the course of his study, Milgram had processed some 780 subjects through twenty-three different permutations of the obedience paradigm, a breath-taking undertaking unmatched ever since. The majority of the experiments were audiotaped. Milgram also recorded hours of conversations of subjects during a debriefing with a psychiatrist. Perry interviewed dozens of former subjects, surviving family members of the actors who played the parts of the "scientist" (John Williams) and the "learner" (Jim McDonough), and analyzed the mountains of documentation and correspondence that Milgram accumulated during the research. She made five important discoveries about the research: (a) there was a great deal of evidence that many subjects were traumatized by their participation, (b) there was a great deal of evidence that many subjects were skeptical about the cover story, (c) Milgram did not follow the protocols for encouraging obedience that he had published, (d) he did not publish all his results, suppressing information that could have jeopardized his overall conclusions, and (e) and he did not debrief the majority of the subjects immediately after the experiment (Brannigan 2013a).

Where experimentalists populate their publications with nameless subjects, Perry exposes the actual individuals who were recruited as subjects. Herb Winer was "boiling with anger" for days after the experiment (Perry 2012:79). At the time, like Milgram, he was an untenured professor at Yale. He confronted Milgram in his office with his concerns about the experiment, particularly about pressure to shock someone with a heart condition. His trauma was so intense that he confided in Perry, nearly 50 years later, that his memory of the event would be "among the last things I will ever forget" (2012:84). After the cover story was explained, Winer became an admirer of Milgram, "although he will never forgive him for what he put him through." Bob Lee was another subject tracked down by Perry – deceived, angry and humiliated. Yet another, Bill

Menold, was unsure of whether the study was a sham or not, but he found it "unbelievably stressful … I was a basket case on the way home" (2012:52). He confided that night in a neighbor who was an electrician to learn more about electrical shocks. Hannah Bergman (a pseudonym) still recalled the experiment vividly after half a century. Her recollections suggested that she "was ashamed – and frightened." Her son told Perry that "it was a traumatic event in her life which opened some unsettling personal issues with no subsequent follow-up" (2012:112). A New Haven Alderman complained to Yale authorities about the study: "I can't remember ever being quite so upset" (2012:132). One subject (#716) checked mortality notices in the *New Haven Register*, for fear of having killed the learner. Another subject (#501) was shaking so much he was not sure he would be able to drive home; according to his wife, on the way home he was shivering in the car and talked incessantly about his intense discomfort until midnight (2012:95). Subject 711 reported that "the experiment left such an effect on me that I spent the night in a cold sweat and nightmares because of fears that I might have killed that man in the chair" (2012:93). None of the previous histories of these experiments even hinted at such reactions, and neither was any of this ever reported in the university curriculum. What caused all the trauma?

To say that the de-hoaxing left a lot to be desired would be a gross understatement. In his first publication, Milgram had written that steps were taken

> to assure that the subject would leave the laboratory in a state of well-being. A friendly reconciliation was arranged between the subject and the victim, and an effort was made to reduce any tensions that arose as a result of the experiment.
>
> (Milgram 1963:374)

Also, "at the very least all subjects were told that the victim had not received dangerous electric shocks." Perry's review of the archives indicates that this was simply not the case. In fact, Perry reports that 75% of the subjects were not immediately debriefed in any serious way until the last four out of twenty-three conditions. Perry reports that around 600 people left the laboratory believing that they had shocked a man, with all that dramatized agony etched on their conscience (2012:92). This was corroborated by Alan Elms, Milgram's research assistant in the first four conditions. "For most people who took part, the immediate debrief did not tell them there were no shocks" (2012:90). In addition, many of the subjects who met after the completion of the study with the psychiatrist, Dr. Paul Errera, similarly reported they received no debriefing at all (2012:89–107). At minimum, a debriefing would have involved an explanation that the scientist and the learner were actors, the shocking appliance was a fake, all the screams were simulated, and that the teachers were the focus of the study. Perry reports that even where some account was given by Milgram to the subjects, they were told that their behaviors, whether obedient or defiant, were natural and understandable, and that the shocking device had been developed to test small animals and was harmless to people. So, even when it occurred, the debriefing, in Perry's words, "turned out to be another fiction" (2012:90). In addition, the debriefing was remarkably brief – two minutes – and did not involve any question–answer interaction with the experimenter. Milgram did not want future subjects to be contaminated by accounts from prior subjects about the true nature of the experiment, and so he withheld such information until the experiment was virtually over. A fuller explanation was mailed to subjects a year later, but it does not seem to have consoled any of those interviewed by Perry.

If many subjects were traumatized, there were significant others who had their doubts about the cover story (2012:156). One subject wrote to Professor Milgram the day after his participation. He had inferred that the "draw" for roles was fixed, and that both pieces of paper probably had the word "teacher" written on them. He found the learner unaccountably "disinterested" and was suspicious of all the one-way glass mirrors. He also noticed that the learner was not given his cheque at the same time as himself. Another noticed that the learner's cheque was dog-eared from what appeared to be frequent use. Others engaged in reality testing by asking the learner to tap on the wall if he could hear him. No response. One *lowered* the shock level intentionally, and the learner seemed to express *increased* pain despite this. Others were simply skeptical that Yale would permit anyone to absorb such punishment. Some commented on the fact that no one with a cardiac condition which was under medical surveillance would submit to such intense agitation. Another noted that there was a speaker in the learner's room, and the sound from the voice did not appear to be coming through the door, as he would have expected. And many suggested that the sounds appeared to be audio recordings. All this was noted in the archives. Under these conditions, the subjects simply played along as required by the experiment, since they assumed that no one would purposely be hurt, and it was all for the good of science.

Milgram was aware of this skepticism, but he dismissed it as a reaction formation. He reasoned that the subjects had acted shamefully, then, in self-defence, they denied anyone was injured, and that they had not done any harm. Perry (2012:163) comments further that "only half of the people who undertook the experiment fully believed it was real, and of those, two-thirds disobeyed the experimenter". There was another area of information leakage that must have piqued the curiosity of some teachers. There were numerous cases where the subjects practically shouted out the correct answer to the learner, but this communication never made a difference in his response. Also, numerous teachers, frustrated by the learner's poor performance, offered to switch places during the experiment, but again, this offer did not attract any interest or response. This did not always result in outright disbelief, but created some suspicions that things were not exactly as they seemed.

Among the unpublished investigations, Perry discovered a remarkable condition that Milgram had kept secret. This was the study of "intimate relationships." Twenty pairs of people were recruited on the basis of a pre-existing intimacy. They were family members, fathers and sons, brothers-in-law, and good friends. One was randomly assigned to the teacher role, the other to the learner role. After the learners were strapped into the restraining device, Milgram privately explained the ruse to them, and encouraged them to vocalize along the lines employed by the actor in response to the shocks in previous conditions. The "intimate relationships" study produced one of the highest levels of defiance of any condition: 85%. It also produced a great deal of agitation in teachers as the learners begged their friends or family members *by name* to be released. One subject (#2435) lost his composure with the scientist's pressure and started shouting at him for encouraging him to injure his own son.

Perry speculated that Milgram was ambivalent about this condition for two reasons. On the one hand, "Milgram might have kept it secret because he realized that what he asked subjects to do in Condition 24 might be difficult to defend" (2012:202). After all, he abused their mutual trust and intimacy to turn the one against the other. On the other hand, the results countered the whole direction of Milgram's argument about the power of bureaucracy. Perry found a note in the archives in which Milgram confessed that "within the context of this experiment, this is as powerful a demonstration of disobedience that can be found." When people believed that someone was being hurt, and that it was someone close to them, "they refused to continue" (2012:202). Given its implication, the finding was never reported.

This suggests that, to an extent, Milgram cherry-picked his results for impact. Perry notes that Milgram worked to produce the astonishingly high compliance rate of 65%. He assumed that he needed a plurality of his subjects, but not a figure so high that it begged credibility. As Russell (2011) noted, in pilot studies he tweaked the design repeatedly. Milgram explored a number of Stress Reducing Mechanisms and Binding Factors to optimize compliance. Stress was reduced, for example, by framing the actions as part of a legitimate learning experiment, and by advising the subjects that there was no permanent damage from the shocks. The binding factors included the gradual thirty-step increments from the lowest to the highest shock level on the supposition that once they started, the movement up the shock scale would signal their acceptance of the protocol one step at a time.

Perry also found that there was often a Mexican standoff between the subjects and Mr. Williams as to their point of defiance. This was particularly evident in the all-women design. In their histories of the experiments Blass (2004 and Miller (1986) created the impression that the scientist would use four specific prods to encourage the subjects to continue, since that was what Milgram published. "If the Subject still refused after this last [fourth] prod, the experiment was discontinued" (Blass 2004:85). The subjects were always free to break off. After listening to the All-women Condition (condition 20), Perry concluded: "this isn't what the tapes showed" (2012:136). Mr. Williams did not adhere strictly to the protocol.

In his face-to-face dealing with subjects, Milgram assured them that their reactions were normal and understandable. Yet, in his book, he describes the compliant subjects as acting in "a shockingly immoral way" (1974:194). In his notes, he describes them as "moral imbeciles" capable of staffing "death camps" (Perry 2012:260). In the 1974 coverage of his book on the CBS network *60 minutes* program, he portrays the compliant subjects as New Haven Nazis (2012:369), and asserts that one would be able to staff a system of death camps in America with enough people recruited from medium-sized American towns.

Replicating Milgram

Having noted these issues in Milgram's original work, we also have to acknowledge attempts at replication. J. M. Burger (2009) replicated Milgram's work, at least partially. His work was based on a revised approach in which the learner reports medical problems with his heart, and the teacher receives remote voice feedback from shocks appearing to originate in a separate room. Given the worries over the potential traumatization of subjects caused in part by Milgram's original work, Burger limited the maximum shock level to 150 volts. In Milgram's original study, 79% of persons who went *beyond* this level of shock showed total obedience. This was also the point at which the learner first expressed serious complaints, and loudly demanded to be released from the study. Burger measured whether the subjects *tried* to continue after the 150-shock level; all subjects who had not desisted at this point were prevented from continuing. The experimenter avoided the prolonged pressure on the subjects to comply at higher shock levels, while getting a measure of aggression that correlated with the original conditions and findings.

The most interesting finding from Burger's replication was reported in a second paper in which he analysed responses to the prods from the "scientist". Burger argues that Milgram was not really studying obedience to orders at all. In the original study, if a teacher hesitated after resistance from the learner, the "scientist" used four escalating prods to get him or her to continue: "please continue", "the experiment requires that you continue", "it is absolutely essential that you continue", and "you have no other choice, you must go on". Only the last prod looks anything like an order. "When participants heard the only prod that we might reasonably consider an order, not one individual 'obeyed'" (Burger, Girgis, and Manning 2011). Indeed, the evidence shows that compliance declined with each level of escalation. Burger et al. concluded by noting that alternative interpretations to Milgram's work should be explored and "the way the research is portrayed to students, scholars, and the public may need to be reassessed" (2011:6). The inconsistencies in the way in which the prods were employed were also identified in Gibson's analysis of the archives (Gibson, 2013a).

Conclusion

Almost six decades after Milgram explored obedience, his work continues to attract attention as the most provocative experimental research in social psychology. However, we have powerful and compelling evidence that, taken as a whole, most people did not behave like Nazis. Where they were convinced of the painfulness of the shocks, they tended to defy pressure to obey. And when they complied they didn't think anyone was actually being harmed. Also, even where replicated by Burger, the behavior examined in the laboratory is not obedience as such.

Zimbardo and the Stanford prison experiment (SPE)

In the early 1970s, Philip Zimbardo created a simulated prison at Stanford University's Department of Psychology. He screened seventy potential volunteers from the Stanford University student body before selecting "about two dozen young men" (Zimbardo 1972) who were randomly assigned to roles of guards and inmates in a makeshift prison. They were paid fifteen dollars per day, although the guards served only an eight-hour shift while the inmates were detained twenty-four hours a day.

> The inmates were unexpectedly picked up at their homes by a city policeman in a squad car, searched, handcuffed, fingerprinted, booked at the Palo Alto station house and taken blindfolded to our jail. There they were stripped, deloused, put into dress-like uniforms, given a number and put into a cell with two other inmates where they expected to live for the next two weeks.
>
> (Zimbardo 1972:4)

By the fourth day, three inmates were dropped from the experiment due to "acute situational traumatic reactions such as crying, confusion in thinking and severe depression" (1972:4). Five out of the eleven "inmates" would eventually leave the study prematurely because of trauma. Many of the guards began acting with cruelty and brutality toward the mock inmates. Although he reported that what he saw was "frightening," Zimbardo let this go on for another three days, filming some of the behavior for television news, before cancelling the experiment (Haney, Banks, and Zimbardo 1973; Zimbardo 2007). Zimbardo's justification for ending the experiment was not as altruistic as one would have imagined but, rather, self-centered:

> in the end, I called off the experiment not because of what I saw out there in the prison yard, but because of the horror of realizing that *I* could have easily traded places with the most brutal guard or become the weakest prisoner.
>
> (Zimbardo 1972:6)

His later charge that the Institutional Review Boards "overreacted" to this sort of abuse of subjects is highly questionable. During this period, where was the American Psychological Association and its ethical standards? One wonders whether the provocative detention of the subjects by the Palo Alto police raised similar questions among civil libertarians. The study also raises questions about whether citizens can "voluntarily" agree to suspend their rights to physical security without knowing that they will be stripped naked in front of strangers and sprayed with deodorant, and asked to dress in female frocks without underwear for two weeks. Leon Festinger (1980:251–52) had this comment,

> From my biased point of view, there was some confusion between "relevant" and "newsworthy." Certainly, if some finding was picked up by the mass media, that was clear evidence that it was relevant. One can improvise a jail and have subjects volunteer … One can then report some interesting reactions of certain individuals. It's an important topic and clearly newsworthy. But it's not research, does not seriously attempt to look at relationships between variables, and yields no new knowledge. It's just staging a "happening."

Festinger's point is that the prison simulation study was merely staging a happening, more like guerrilla theatre than serious science. There was no hypothesis identified in advance, except the idea that people who were of average or normal backgrounds will take on situational roles no matter how much for the worse. "The mere act of assigning labels to people and putting them into a situation where those labels acquire validity and meaning is sufficient to elicit pathological behavior" (Zimbardo 1972:6). The research was a fishing expedition, designed less to test relationships between variables and to advance theory than a device to dramatize the supposedly well-known proclivity of prison guards to treat their inmates with inhumanity. It suggested that the deplorable misconduct of some guards in some institutions arises from the situational roles of dominance and subordination, and not from individual traits. If true, this would be quite a breath-taking inference, and would assign role theory pride of place in the theoretical arsenal. Yet, Zimbardo reports significant variation in the posture of the guards and reports that half the inmates had to be dropped from the experiment prematurely. Zimbardo's perspective makes a virtue out of the experiment's inability to tap the stability of traits over the life course and across different contexts by implying that the frictions of prison life reflect the largely situational conditions that short-term experiments can turn on and off at will.

I do not quite know what to make of the Zimbardo study. If it was as traumatizing as he alleges, I cannot understand why he was not sued, dismissed, and/or censured for ethical misconduct by the APA. On the other side, if the students were merely pretending to be cruel, that is, acting, then Zimbardo's conduct is less culpable, but his conclusions are less relevant. The other possibility is that subjects were invited to role-play in a situation where the "play" was *not* simulated. The inmates did not *pretend* to strip naked, and wear "dresses" and ankle shackles 24 hours a day or dress without undergarments. This actually happened. Zimbardo designed a situation that was intended to humiliate the subjects – and a third departed the study within days. Likewise, the guards did not carry toys, but real wooden nightsticks. So the subjects were put in a highly ambiguous field of play where action drifted back and forth across an experiential border demarcating heartfelt impulses and mere acting. Consider the account from a guard-subject:

> During the inspection I went to cell 2 to mess up a bed which the prisoner had made and he grabbed me, screaming that he had just made it, and he was not going to let me mess it up. He grabbed my throat, and although he was laughing I was pretty scared. I lashed out with my stick and hit him in the chin (although not very hard) and when I freed myself, I became angry.
>
> (Haney et al. 1973:88)

Does this mean that the guard and prisoner lost control, began to act "in earnest," and that the subject was consequently assaulted? Another ambiguity in the situation may have arisen from the fact that the experimenters themselves appear to have been swept up in the play and lost their scientific detachment. "Over time, the experimenters became more personally involved in the transaction and were not as distant and objective as they should have been" (Haney et al. 1973:78) In fact, Zimbardo played the role of superintendent. Did the superintendent carefully censor every outburst of his guards and act as a model of virtue? No. Yet, he would later characterize *their* behavior as "aberrant, anti-social behavior" (1973:90) without consideration of whether the subjects took his complicity as approval for their actions. The more one reflects on the shift back and forth between simulation and "spontaneous" aggression, the more apparent it is that the mock prison was not so much an innocent analog of real prisons, but a species of the very reality it was meant to mimic. In that case, the "revelation" that persons who play certain roles actually come to exhibit traits of the persons who perform them for a living is rather shallow, since the "play" here included conditions of degradation that were not mere play, that is, actually stripping in front of strangers, dressing in demeaning clothes, sleeping in ankle shackles, being wakened from (real) sleep in the middle of the night on the pretext of a count, etc. None of this was simulated. In this reading, subjects were not just tested; they were humiliated. And the experimenter fired off an article to a journal (*Society*) to condemn the sort of behavior in prison officials that he had himself created during his not-so-mock exploration of the same subject.

Either way, Zimbardo's work illustrates a point from the previous chapter – much research is simply ethics in disguise, in this case, a telling criticism of prison life. Festinger's dismissive summary seems to imply that Zimbardo's use of the experiment as a stage is wholly at variance with the field. My view is that this use of the experimental idiom as a stage to dramatize something was actually more common in the classical period than people allow, however little attention is paid to this practice in methodology courses and textbooks (Griggs and Whitehead 2015). Indeed, a great deal of the work that appears to devolve from formal theory testing is, on the contrary, quite serendipitous.

The archival re-evaluation of the SPE by Thibault Le Texier

Recently, Thibault Le Texier (2018, 2019) employed archival materials to piece together how the SPE was designed and implemented. He contacted fourteen participants from the original experiment for interviews by telephone. He also reviewed all of Zimbardo's audio and video recordings, his publications, and blogs. The audio and video materials have been transcribed, and most printed material has been digitized, permitting research online. Le Texier's resulting account, "the story of a lie," does not mince words. Le Texier outlines how Zimbardo produced a standardized account of the experiment in a slide show which he used for decades to promote the study. The slide show was also presented widely to military groups and may have had a role in training US military interrogators in Iraq and Afghanistan (Le Texier 2018:45–47). What the standardized account leaves out is that the SPE was largely based on a project by students in Zimbardo's social psychology class, undertaken in the spring of 1971, three months before Zimbardo launched his own study in August. In a term paper titled "A Simulated Prison", dating to spring of 1971 which was found in the archive, David Jaffe (1971) outlines how he and other students designed a class project to be run over a single weekend on campus in the Toyon Hall residence at Stanford to simulate the harmful psychological effects of imprisonment.

> [W]e derived three basic goals, or psychological effects we intended to produce in our inmates. First, we wanted to create the loss of freedom. For the entire weekend, where inmates were, what they did, how they did it, etc. were not to be under their control, but rather under the control of the prison staff … Second, we wanted inmates to feel they depended on the prison staff for the satisfaction of all their needs, even those as basic as food and toilet privileges. Finally, we decided to try to produce a feeling of deindividuation, partly by forcing the guards to deal with the inmates in groups, and partly by costuming the inmates in ugly, standard prison gowns.

Le Texier documents the extent to which Zimbardo borrowed from the student study. He lays out side by side the rules which the students had created in the Toyon Hall study with those which he claims the guards in the SPE came up with on their own. Zimbardo also reported to Leslie Stahl on a *60 Minutes* report (August 30, 1998) that he did not create the rules himself, that they were set by the guards (see Le Texier 2018:62). As Le Texier says: "Instead of recognizing the foundational importance of Toyon Hall's experience, Zimbardo completely obscured it for forty years" (2018:61–2).

Le Texier argues further that the subjects recruited as guards were instructed on the day before contact with the inmates as to how they were expected to act by assuming a domineering and aggressive posture. During the study, Zimbardo, as the prison administrator, also instructed the guards to be sarcastic and to humiliate the inmates by arbitrarily depriving them of privileges. Zimbardo's student warden encouraged the guards to use their whistles during the 2:30 am cell count to irritate them and awaken them violently. The guards, while paid as experimental participants, were encouraged to think of themselves as research assistants, as helping agents, and not as subjects of study themselves.

As for the evidence of trauma among the inmates, this was based in part on the necessity of early release of some inmates. However, in one case, the individual reported afterwards that he faked extreme emotional discomfort to get out of the experiment. It turns out that the inmates were not entitled to watch TV and read, as they had been promised, and some were bored. Also, *post hoc* questions of the guards suggests that the majority of them (Le Texier 2018:131) never completely got into their roles, and were always aware that they were in a simulated environment. The same is reported from interviews with the inmates. Zimbardo did not collect the sort of thorough conversational exchanges found in Milgram, so it is impossible to examine systematically the actual course of guard–inmate interaction. However, it is clear that Zimbardo studiously blocked attempts of subjects to depart from the experiment when they tried to do so, a breach of ethics at least as questionable as anything in the obedience studies. However, this did not attract the same scrutiny that the National Science Foundation applied to Milgram, since his funds were provided by the U.S. Office of Naval Research. It is also clear that the critical conclusions about the apparently deleterious nature of prisons was a foregone conclusion. Although several of these points have been in the literature for decades, Le Texier's originality is in tracking down the archival evidence that has forced academics to distance themselves from the scientific value of the SPE (see Zimbardo 2018).

Rosenhan on being sane in insane places

The last vintage demonstration "experiment" to be discussed was undertaken by David Rosenhan (1973), "Being Sane in Insane Places." In this study, eight healthy volunteers faked symptoms of mental illness and were incarcerated in twelve different hospitals in five states to determine whether psychiatrists could distinguish insanity from normality. They all faked an auditory hallucination. They told psychiatrists that they heard voices say "empty", "hollow", and "thud". Once admitted they were instructed to act normal, and to avoid taking the psychiatric medicine. When they spit it into the toilets, they noticed that many of the real patients were doing the same thing. All the pseudo-patients, except one, were diagnosed with schizophrenia. They were retained for a period of from eight to fifty-two days. The study was published in *Science*, the most prestigious scientific publication of the day. This study was cited as evidence for the environmental theory of mental illness by suggesting that maladaptive social conduct is a function of oppressive and dehumanizing medical institutions. And, certainly, Rosenhan's reports from the inmates supported the view of callous conditions of treatment. However, evidence of callous treatment in total institutions was already well-known to Goffman (1961). What Rosenhan's supporters failed to acknowledge sufficiently was that the normal volunteers acted insane to gain admission. The study was a landmark "demonstration" of the oftentimes inhuman effects of total institutions, but it was not a deductive experiment, however much attention it attracted, and neither did it advance our theories of institutions in any significant way. It did not explain what type of facilities were more likely to be dehumanizing and neither did it throw any light on the factors that lead to dehumanization (cultural beliefs, individual differences, hospital policies, etc.). But it enjoyed enormous appeal because of the moral subtext – people already suffering from a mental handicap were brutalized, if the reports were accurate and representative, by the agents created to ameliorate their discomfort.

The credibility of Rosenhan's research has been questioned in a recent book by Susannah Cahalan (2019). Cahalan was able to track down two of Rosenhan's pseudo-patients – Bill Underwood and Harvey Londo. Underwood reported a lot of brutal and de-personalizing experiences in his hospital stay, but Harvey Lando had quite the opposite experience. He was held for nineteen days in the San Francisco public health facility and found that the psychiatric professionals were extremely supportive of him. He subsequently became a professor of psychology at the University of Minnesota, but his conclusions were so contrary to the thesis that Rosenhan was intent on establishing that he was dropped from the study and relegated to a footnote. The other informant was Rosenhan himself. Cahalan employed a private investigator, but could not find evidence for the existence of the other six pseudo-patients that Rosenhan claims were involved. They were also unknown to Rosenhan's research assistant at the time. An advertisement in *Lancet Psychiatry* failed to produce any contacts. Cahalan was forced to conclude that Rosenhan simply made them up.

She also recovered some of the admission documents associated with Rosenhan's own experience. Rosenhan was held for nine days. Cahalan found that, in addition to the standard auditory hallucinations which all the pseudo-patients were supposed to report, he reported that he had suicidal tendencies, as well as ringing in his ears that he had to cover with copper covers and that he was committed to treatment by his wife. After his work was published in *Science*, the Haverford State Hospital in which he was treated leaked his file to Robert Spitzer, the editor of the *Diagnostic and Statistical Manual of Mental Disorders*. He appears to have suspected fraud in Rosenhan's report. We will never know for sure. What we do know is that after being paid a large financial advance and writing 200 pages of a book on his "experiment", the project was simply dropped and Rosenhan never returned to it. He was, according to colleague Lee Ross, extremely secretive about his work. There is nothing in his archives substantiating the identity or contribution of his six unidentified collaborators. He was, in the view of Susannah Cahalan, “the great pretender”. However, his work contributed significantly to the demonization of psychiatric institutions, and the movement towards the de-institutionalization of the mentally ill in the 1970s. That change contributed directly to another social problem – urban homelessness.

The studies which we have reviewed in this chapter and in the previous one are landmark accomplishments in experimental social psychology. Although vastly different in duration, location, and subject pools, these provocative and intriguing investigations each attracted academic and public attention because of the importance of the subjects they examined, and because of what they sought to learn from them. In retrospect, the evidence suggests that, to one degree or another, their findings were foregone conclusions and their experimental formats were essentially methods of demonstration: that is, unscientific use of experimental methodology to explore and elaborate moral questions.

5 Bystander research

Plumbing the psyche of the indifferent Samaritan

Introduction

Social influence has been one of most fruitful areas of research in classical social psychology. Sherif's analysis of adolescent peers in the Robbers Cave field experiment sought to show how adversity molded strangers into cohesive groups and expedited inter-group conflict. Asch's work suggested how *ad hoc* groups could influence individual perceptions in the line discrimination experiments. And, in some of his experimental variants, Milgram showed how teams of actors could mediate the power of malevolent authority figures to reduce obedience (Milgram 1965). An obvious extension of this literature has been a concern for the irrational influence of crowds on individual behavior, and the associated consequences in terms of rioting and looting (Le Bon 1895). In all these cases, the presupposition has been that social collectivities, whether adolescent peers, crowds, or colleagues, expedite individual behavior and pressure it to develop in certain ways. In short, groups promote individual action.

In the late 1960s, experimentalists turned their attention to the related problem of social influence in a more diffuse situation – that of persons whose salient point of commonality is that they are all observers of some provocative event, such as an emergency, a crime, an accident, or a natural disaster. These types of "associations" promote withdrawal. Moreover, this kind of situation has an implied moral subtext because it raises an implicit duty to respond or intervene to rescue individuals at risk of harm. By contrast, groups of observers in the viewing stands at an airshow or at the lookout on top of Niagara Falls may share a common experience of exhilaration, but no one would employ the term "bystanders" to describe their situations. The concept of bystander implies that the individual is disengaged, or even inhibited, from action. The bystander research was designed *to analyze the conditions under which observers become bystanders*, and how social conditions that imply obligations on individuals to act have the effect, ironically, of making them passive. The parable of the Good Samaritan in the Gospel of Luke tells the story of a traveller left to perish on the roadway after being beaten and robbed. His plight was observed by a priest and a Levite, both of whom avoided him. The Samaritan, even without any obligation to provide assistance, did so. The Bible made his altruism a template for exemplary behavior. For psychologists, the empirical question was what conditions lead persons to become bad Samaritans. The event that triggered the research was a notorious crime in New York City in 1964 – the rape and murder of Catherine ("Kitty") Genovese. Reporter Martin Gansberg reported the event in a story on page one of the *New York Times* on March 27th, under the provocative, four-column headline:

37 WHO SAW MURDER DIDN'T CALL THE POLICE

Apathy at stabbing of Queens Woman Shocks Inspector

> For more than half an hour thirty-eight respectable, law-abiding citizens in Queens watched a killer stalk and stab a woman in three separate attacks in Kew Gardens. Twice, the sound of their voices and the sudden glow of their bedroom lights interrupted him and frightened him off. Each time he returned, sought her out and stabbed her again. Not one person telephoned the police during the assault; one witness called after the woman was dead.
> (Gansberg 1964:1)

Catherine Genovese was an attractive twenty-eight-year-old daughter of Italian immigrants living with her girlfriend in Kew Gardens. She was murdered within steps of her second-floor Tudor apartment after 3:00 am on March 13th following her late-night shift as a bartender at Ev's 11th Hour bar in Queens. On March 18th, police apprehended a suspect on a charge of burglary. He subsequently confessed to several rapes and murders, including that of Kitty Genovese. His name was Winston Moseley, a twenty-nine-year-old African American. He was married, owned a house, and was gainfully employed as a punch card technician at Raygram, a business machine company. He had no criminal record. The first attack on Genovese occurred outside in front of the high-rise Mowbray apartment facing Genovese's apartment. This apartment building gave the inhabitants front-row seats to a brutal attack, or so it was said at the time. The story of the thirty-eight indifferent observers touched a public nerve and attracted unprecedented comments from experts and politicians alike about urban apathy and the decline of community. Ironically, Kew Gardens had been regarded as a safe, crime-free neighborhood of Queens.

Understanding indifference: the diffusion of responsibility among bystanders

The first experimental study of bystander behavior was reported in 1968 at New York University and was modeled explicitly by John Darley and Bibb Latané on information from *Thirty-eight Witnesses*, a book written by the *New York Times* city editor, A. M. Rosenthal in 1964. Rosenthal's account became the canonical narrative based on witness inaction and was adopted enthusiastically by Darley and Latané.

> At least 38 witnesses had observed the attack and none had even attempted to intervene. Although the attacker took more than half an hour to kill Kitty Genovese, not one of the 38 people who watched from the safety of their own apartments came out to assist her. Not one even lifted the telephone to call the police.
>
> (Darley and Latané 1968:377)

They also wrote that "each observer, by seeing lights and figures in other apartment house windows, knew that others were also watching" (1968:377). If one assumes one is the lone observer, the onus is unavoidably on one to respond. However, if the observer knows that he or she is only one of many, "the responsibility for intervention is shared among all the observers, and is not unique to anyone. As a result no one helps" (Darley and Latané 1968:378). From their perspective, observer inaction of the type described by Rosenthal arose from the diffusion of responsibility. To test this idea, Darley and Latané devised an experiment. "Fifty-nine female and thirteen male students in introductory psychology courses … were contacted to take part in an unspecified experiment as part of a class requirement" (1968:378). During the course of this exercise, the subjects were exposed to an emergency which was simulated. The subjects were told that they were recruited to discuss adjustment problems at large urban universities. In the interests of anonymity and to prevent potential embarrassment, each participant would communicate with the others through a microphone and headsets from individual cubicles. The experimenter explained the protocol for introducing problems, suggesting how the participants would assess them and come to a collective solution. Following these discussions, the experimenter apparently went "off air" to permit the subjects to proceed on their own initiatives each turn-taking in two-minute sequences. In addition, the subjects were led to believe that when any person spoke, the other microphones were turned off. Only one person could be heard at a time. The emergency which the experimenters introduced was what sounded like an epileptic seizure of one of the participants, characterized by choking, stuttering and calls for help indicating that the person was in grave distress. The experimenter varied the number of apparent interlocutors, but there was only one real subject. The utterances of the other participants were all pre-recorded. The major independent variables were (1) the group size and (2) the differences in gender composition and expertise of the group. The dependent measure was the time elapsed between the onset of the victim's fit and the subjects' departure from the cubicle to seek assistance. The experiment was terminated after six minutes of the onset of the emergency.

The experimenters created groups of three sizes:

- (the real subject and the seizure victim),
- (the real subject, the victim, and another person), and
- (the real subject, the victim, and four others).

There were dramatic differences in the numbers of subjects who terminated the experiment before it ended or was ended by the experimenter at the six-minute mark. Where the subjects thought they were the sole witnesses of the seizure, 85% terminated to seek help. In contrast those who thought they were in a group of six resulted in only 31% who terminated to seek help. The loners also acted much quicker – in an average of fifty-two seconds versus 166 seconds for the group of six, that is, less than a minute versus nearly two and a half minutes. The differences were statistically significant despite the small cell sizes (Table 5.1).

To test for differences in the effect of gender composition and expertise in the groups, the experimenters used the three-person group protocol but varied the gender composition of the third member of the group (in addition to the real subject and epilepsy victim). They also included a male who said he had experience as a medic in an emergency ward at Bellevue Hospital. In three out of the four conditions reported, the real subjects were female. None of the variations produced any significant differences in either the percentage who terminated or the time to terminate. "Subjects responded equally frequently and fast whether the other bystander was female, male, or medically experienced" (Darley and Latané 1968:381). Afterwards, the subjects completed a number of scales (Machiavellianism, anomie, authoritarianism, social responsibility as well as vital statistics and socioeconomic data). None of the results of these scales appear to have correlated with whether the subjects tried to intervene or not.[1] The main effect was group size.

The experimenters also administered a fifteen-item checklist to tap what went through the subjects' minds when they heard the epileptic fit. There were only three statements which attracted high levels of agreement: "I didn't know what to do" (eighteen out of sixty-five); "I didn't know exactly what was happening" (twenty-six out of sixty-five); and "I thought it must be some sort of fake" (twenty out of sixty-five) (Darley and Latané 1968:381). At the end of the experiment, when subjects reported the apparent trauma of the epileptic participant, or when the assistant retrieved the non-responsive subjects from their cubicles at the six-minute termination point, the experiment's assistant "disclosed the true nature of the experiment, and dealt with any emotions aroused in the subject" (1968:379). It is unclear whether subjects were told that the fit

Table 5.1 Effect of group size on likelihood and speed of response

Group size	*N in group*	*Percentage responding by end of the fit*	*Time in seconds*
2 (Subject & victim)	13	85	52
3 (Subject, victim, and one other)	26	62	93
6 (Subject, victim, and four others)	13	31	166

p value of difference: chi sq. = 7.91, $p < 0.02$.

was, in fact, a simulation. As a result, it is difficult to determine whether the skepticism that was reported by 30% of the subjects (20/65) was based on their experiences during the experiment or something conveyed in the debriefing. The report is unclear on this point. Darley and Latané (1968:381) concluded that "subjects, whether or not they intervened, believed the fit to be genuine and serious." This was based on reactions recorded when the epileptic fit occurred. Many subjects verbalized their concerns. However, the authors also mention that the manipulation or cover story was not always successful. "Almost all the subjects perceived the fit as real" (1968:379). Almost all, but not everybody. In fact, they reported that some data were censored, apparently because of subject skepticism. "There were two exceptions in different experimental conditions, and the data for these subjects were dropped from the analysis" (1968:379).

A second major test of the diffusion of responsibility theory was also published in 1968 by Latané and Darley and was conducted at Columbia University. In this design, the experimenters focused on a different kind of emergency, and examined the role played by direct contact with other witnesses exposed to the same situation. Where the previous study at New York University used male and female students in an introductory psychology class, this study recruited male students in graduate programs, in professional faculties, as well as male undergraduates. The students were asked to participate in an interview about student problems in attending classes in an urban university. When they arrived, the subjects were asked to fill out a series of questionnaires in a waiting room, ostensibly before the interview was to occur. The room had an observational "one-way glass mirror" which permitted subjects to be secretly monitored. The emergency event was introduced after the prospective subjects had sat down to fill out their questionnaires. After a couple of minutes, smoke suddenly began to pour out of a wall vent. There were three treatment conditions. First, subjects were exposed to the mysterious smoke while sitting alone. Second, subjects were sitting in a room with two confederates who also were ostensibly waiting for the interview. They ignored the smoke, shrugged their shoulders and continued to complete the questionnaires. Third, three genuine subjects were exposed simultaneously to the smoke treatment.

In the first condition (n=24), 75% of the subjects departed the premises to report the problem within the six-minute test period. In the second condition, which was run ten times with a real subject and two indifferent confederates, *only once* did the real subject depart the room to report the problem (one in ten, or 10%). And in the third condition, which was run eight times with three real subjects in each test (n=24), only three times did *one* of the real subjects respond to report the problem.

The most responsive subjects were those tested alone (75%), and the least responsive were those in groups where the confederates behaved indifferently (10%). Groups with three individuals showed a middling response (37.5%). This was the most puzzling finding. "Only one person reported the smoke within the first four minutes before the room got noticeably unpleasant. Only three people reported the smoke within the entire experimental period" (Latané and Darley 1968:218). This seems to suggest that twenty-one subjects were not actually observing an emergency as much as they were experiencing it. How serious was the smoke? "For the entire experimental period, the smoke continued to jet into the room in irregular puffs. By the end of the experimental period, vision was obscured by the amount of smoke present" (Latané and Darley 1968:217). In the second condition, Latané and Darley describe the situation this way:

> of 10 people run in this condition, only 1 reported the smoke. The other nine stayed in the waiting room as it filled up with smoke, doggedly working on their questionnaire and waving the fumes away from their faces. They coughed, rubbed their eyes, and opened the window.
>
> (1968:217–18)

In a later publication, Latané and Darley (1970:46) described the situation this way: "By the end of four minutes enough smoke had filtered into the room to obscure vision, produce a mildly acrid odour, and interfere with breathing."

Unlike the earlier experiment, with the epileptic fit, the subjects in this experiment were not so much bystanders in the emergency as victims themselves. The authors appear to acknowledge that this was quite a different situation than the previous design, and required a different explanation. "The diffusion explanation does not fit the present situation" (1970:220–21). Alternatively, what they point to is a process of social comparison following the work of Stanley Schachter (1959) in which individuals in stressful situations take cues from others in the same situation to assess how much danger they are actually experiencing. In the debriefing, the subjects reported that they did not react with alarm because "they decided that there was no fire at all and the smoke was caused by something else" (Latané and Darley 1970:220). The smoke did not smell of combustion because it was actually an inert gas, titanium dioxide, surreptitiously released from a bottle by the experimenters through the vent.[2] It defies common sense that five groups of subjects would sit through six minutes of such aversive exposure when individuals attending alone bolted out of the room almost immediately. It also beggars the imagination as to why, in each of the five non-responsive groups, consensus was *always* that the smoke was harmless and that their behavior was, in the words of Latané and Darley, "reasonable under the circumstances" (1970:220) How reasonable is it when subjects were coughing and rubbing their eyes, and having trouble breathing? If this account is valid, that is, that the subjects' behavior was "reasonable", then it suggests another paradox: maybe the smoke was not an emergency at all. It is regrettable that the authors of this paper do not report any evidence of the credibility of their manipulation, as was done in the previous publication. In that experiment, some subjects were dropped because of their skepticism, and a large number of those who remained suggested that the epileptic fit may have been a fake. The questions of experimental credibility and subject incredulity are more pertinent in the second experiment.

Aside from these issues of internal validity, the paper broke new ground in suggesting a plausible sequence for mobilization in the face of disaster. The first step was that persons had to *notice* the event, and tag it as somehow requiring further reflection. Then they had to develop an *interpretation* that labeled it as disastrous, dangerous and/or extraordinary, and then they had to assume *responsibility* to act.[3] Also, friends were more likely than strangers to act quickly in an emergency (Latané and Rodin 1969). The extrapolation back to the Genovese case which motivated the research was mixed. The diffusion of responsibility illustrated by the first experiment seemed to correspond well to the official account, as published by Gansberg in 1964. Many witnesses looked on and assumed others would take responsibility. But the social comparison element raised in the smoke experiment was irrelevant, since the observers did not have opportunities to compare notes in one another's presence. But this meant that the bystander question was beginning to take on a theoretical life of its own, unconfined to the event from which the initial ideas originated and employing insights from other perspectives in the field. Then the story met a bump on the road that made many observers fundamentally revise their view of the bystander situation.

The epiphany of the indifferent Samaritan

In 2007, three British psychologists, Manning, Levine, and Collins (2007), dramatically changed the narrative around the Genovese sexual assault and murder, and the derivative field of bystander research. Investigations of events associated with the crime and the way it was reported explained how the original narrative was shaped. The *New York Times*' story with which this chapter began was actually the second newspaper coverage of the crime. Genovese's case was just one of 636 murders in the city in 1964, and it was reported initially as just another homicide in the back pages of the *Times* ("QUEENS WOMAN IS STABBED TO DEATH IN FRONT OF HOME") and in the *Daily News* ("QUEENS BARMAID STABBED, DIES") (Maeder 2017). Ten days after the murder, the *Times* Metropolitan Editor, A. M. Rosenthal had lunch with New York City's Police Commissioner, Michael Murphy. Rosenthal was piqued by the fact that two men had confessed independently to the murder of Annie Mae Johnson. That was not what Murphy wanted to talk about. "Murphy said that what struck him about it was not the crime itself but the behavior of thirty-eight witnesses. Over a grisly half hour of stabbing and screaming, Murphy said, none of them had called the police" (Lemann 2014). That was the point at which Rosenthal knew he had a new angle on a provocative story. He dispatched Gansberg with a police detective to re-visit Kew Gardens to investigate the event more fully. This led to the sensational coverage that provoked reaction all over the country. One of the key lines from the many persons interviewed was "I didn't want to get involved." The crime moved Rosenthal in an almost primordial way. In his 1964 book, *Thirty-Eight Witnesses*, he wrote: "there is in the tale of Catherine Genovese a revelation about the human condition so appalling to contemplate that only good can come from forcing oneself to confront the truth" (Rosenthal 1964).

One of the key informants consulted by Manning, Levine, and Collins was Joseph De May, who moved to Kew Gardens after the infamous crime, and who developed an online history of the neighborhood – including the Genovese case. De May, a lawyer, was interested in the evidence about the crime which was gathered by the police for the prosecution of Winston Moseley (De May 2006). The original account published by the *Times* and based on the interpretation of the Chief of Police was misleading in several instances:

> Not all of the 38 witnesses were eye witnesses (some only heard the attack); witnesses have since claimed that the police were called immediately after the first attack; none of the eye witnesses could have watched Kitty or her attacker for the full 30 minutes because they were visible to the witnesses for only a few moments; there were two separate attacks not three; the second attack occurred inside part of a building where only a small number of potential witnesses could have seen it; Kitty was still alive when the police arrived at the scene.
>
> (Manning, Levine, and Collins 2007:557)

At the trial five witnesses were called. The assistant district attorney said that there were only half a dozen witnesses who saw the event. Also, the reason that there were two attacks

> was that Robert Mozer, far from being a "silent witness," yelled at Moseley when he heard Genovese's screams and [Moseley] drove away. Two people called the police. When the ambulance arrived at the scene – precisely because the neighbors had called for help – Genovese, still alive, lay in the arms of a neighbor named Sophia Farrar, who had courageously left her apartment to go to the crime scene, even though she had no way of knowing that the murderer had fled.
>
> (Lemann 2014)

Lemann also reports that there were two other witnesses. One man, Joseph Fink, was the lift operator in the Mowbray apartment building across the street from where Genovese lived. He saw the first attack but went back to his apartment and took a nap rather than go outside to offer assistance. After the first attack, Genovese got to her feet and slowly staggered back to her apartment. Moseley left the scene and, according to court documents, pulled his car into a darker area where his license plate would be less conspicuous. When he returned, he found Genovese at the vestibule of one of her neighbors, Karl Ross. This was when she was sexually assaulted and fatally stabbed. Ross apparently opened his door and witnessed the stabbing.

> He made a couple of phone calls, the first to a friend in Long Island, the second to a neighbor in the building, who told him to come over. Ross crawled out of his window, across the roof, and into the neighbor's apartment, and eventually called the police.
>
> (Lemann 2014)

Ross was frequently drunk and may have been apprehensive about calling the police because he was gay, and the police were notoriously homophobic. Neither Fink nor Ross were called as witnesses. Since Moseley confessed to the charges, the utilization of the witnesses was to decide whether Moseley could be found not guilty by reason of insanity and whether his crime merited the death penalty. He was sentenced to life imprisonment.

The picture that emerges from a sober second take on the original crime could not be further from that which captured public interest in the first instance. There was no large shadowy figure of indifferent observers incapable of acting due to the diffusion of responsibility. In fact, only a few people actually were in a position to witness the crime. Of those who were, one shouted loudly to scare off the attacker and several acted to call the police and an ambulance. Unfortunately, Genovese died in the ambulance en route to the hospital. But the original story, according to Manning, Levine, and Collins, became a kind of modern parable. "Whereas the good Samaritan parable venerates the individual who helps while others walk by, the story of the 38 witnesses in psychology tells of the malign influence of others to overwhelm the will of the individual" (2007:555).

The moral tale is that group observers of emergencies become a malign influence that undermines normal individual altruism. Moreover, the repeated telling of the story hijacked the social psychology of helping behavior, both in the now-classic bystander research of Darley and Latané, and in the social psychology textbook industry. Ironically, the introductory textbooks furnish a more accurate account (Griggs 2015b). Bystander behavior became defined as withdrawal. In the view of Manning et al. (2007:561): "the study of the possible conditions under which groups can facilitate helping seems to have withered on the vine."

In 2016, the *New York Times* published two stories that partially corrected the misleading impressions created by the original Gansberg report,[4] particularly questioning the inflated number of witnesses and their alleged apathy (Dunlap 2016; Haberman 2016). They also reproduced a photo of both stories – the first from March 14th buried in one of the back pages, and the second from March 27th on the front page for the purpose of comparison. The retraction was only partial since Dunlap's story ended by saying that although some of the key facts were wrong, its broader conclusion was indisputable: "that city dwellers are capable of stunning indifference to their neighbor's life-and-death plights" (Dunlap 2016). While this may be true at times, it was not a fair conclusion of the revised understanding of what happened in Kew Gardens on Friday March 13, 1964.

Studies of bystander effects in dangerous situations

One of the paradoxes of the study of bystanders in the Latané and Darley experiment where Ss were exposed to smoke in a university waiting room was that many persons did not perceive the event as dangerous. Perhaps the experimenters expected to provoke fear of fire, but anyone familiar with camp-fires or a fireplace at home would not have equated the smoke at Columbia University with combustion. They thought that maybe the smoke was the evaporation of an air conditioner or something equally benign. Later researchers picked less ambiguous situations, and staged them in more naturalistic, outdoor situations. Harari, Harari, and White (1985) simulated a sexual assault on the San Diego State University campus and arranged to have subjects walk on a pathway that was 20–30 feet from the scene of the assault either alone or in groups of three. In the group condition 85% (n = 34) intervened personally to assist; 15% (n = 6) chose not to intervene. In the individual condition 65% (n = 26) intervened and 35% (n = 14) did not (1985:656). "Group bystanders were at least as likely as individuals to intervene when the setting was natural, when the subject's plight was clear, and when group members could see and talk to one another" (1985:657). Since people could talk to one another, this probably facilitated the definition of the situation as an emergency, reducing the diffusion of responsibility. Fischer, Greitemeyer, Pollozek, and Frey (2006) also studied a sexually aggressive scenario in a laboratory setting using professional actors, and created high risk versus low risk conditions by varying the physical stature of the perpetrator. In the low danger condition, the actor "was a skinny male of small stature" (2006:272). "The female actor representing the victim … was a twenty-one-year-old petite female with a fragile physique. Her counterpart on the video sequence with high potential danger was a strong-built, thug-like male" (2006:271).

> The male actor was requested to dominate the conversation and to flirt more heavily as time progressed. During the third and fourth minute, the male actor was instructed to increase sexual insinuation up to a level of unambiguous verbal sexual harassment, while the victim was told to focus on the completion of the experiment, but then, due to the increasing persistence of the male actor, to defend herself verbally and reject the perpetrator and his statements. In the fifth minute, the male perpetrator loudly starts insulting the victim and touching her without her permission.
>
> (2006:271)

She screams and tries to leave the room but is blocked. A few second later the picture goes blank. These people ostensibly were other participants being monitored live with a camera in an adjacent room, and the subject's task was to evaluate their interaction. When the level of harassment of the female was low, bystanders viewing alone were more likely to intervene (50%) than those in a group of two (9%). However, in the high danger scenario described here, persons acting alone or in a group responded at about the same level (44% versus 40%). The apathetic bystander evaporated when the person at risk was highly vulnerable.

Darley and Latané's research appealed broadly to social psychologists and fostered a wave of replications and extensions to better understand bystander behavior. Fischer and eight others (2011) undertook a meta-analysis of all the bystander studies from the 1960s to 2010 to identify the extent to which the apparent bystander indifference is affected by whether the critical situation is a dangerous versus a non-dangerous emergency. They pulled together 105 independent studies with a total of more than 7,700 subjects. The weighted mean effect size of the help-reducing bystander effect was –0.35 "which amounts to a small but moderate effect" and a negative one at that, that is, more bystanders, less help (2011:523). However, when the cases were sorted separately according to the level of danger facing a potential victim, the bystander effect became positive (g = +0.29), meaning the more bystanders, the higher the level of helping behavior (2011:523). Dangerous situations can expedite interventions because they are arousing and stressful, and the costs of apathy are clearly high, so intervention in fact can reduce observer stress by helping the victim. Bystanders can also be a source of physical support in the face of fear and can diffuse the individual risk of intervention. "We expect that [bystander indifference] substantially declines because of the fact that many dangerous emergency situations can only be resolved by a group … additional bystanders could help to overpower a fierce perpetrator" (2011:521). In addition, some emergencies may be so dangerous that they cannot be managed by individuals acting alone and might require the coordination of several persons. In short, dangerous emergencies attenuate the magnitude of bystander apathy because

> (a) increased levels of arousal … [are] experienced especially in high-danger situations, (b) reduced fear [is] based on the expectation that additional bystanders can provide physical support in dangerous emergencies, and (c) the rational expectation that some emergencies can only be resolved by cooperation and coordination between several bystanders.
>
> (2011:521)

Fischer et al. (2011) also report that the findings of bystander apathy in more recent psychological studies have become less prevalent. Articulating the new reality based on more recent studies, Stadler (2019) suggested that "bystander apathy is not the norm." In his re-analysis of Latané and Nida's 1981 review of the literature, Stadler (2008) showed that "the more bystanders there were, the more likely a victim would receive help, at least when the bystanders could not all see each other (like in the Genovese case)." In a unique study using public CCTV camera footage, Philpot et al. (2019) were able to track down 219 cases of individuals who were both victimized and caught on camera.

> Using a unique cross-national video dataset from the United Kingdom, the Netherlands, and South Africa (N = 219), we show that in 9 of 10 public conflicts, at least 1 bystander, but typically several, will do something to help. We record similar likelihoods of intervention across the 3 national contexts, which differ greatly in levels of perceived public safety. Finally, we find that increased bystander presence is related to a greater likelihood that someone will intervene. Taken together these findings allay the widespread fear that bystanders rarely intervene to help.

Philpot et al. do not discount the fact that sometimes observers do not intervene to help, but the norm appears to be quite the opposite. High levels of involvement were found in different national and urban contexts, suggesting that involvement is the norm in actual public conflicts as captured by the CCTV technology. In addition, the emergencies, conflicts and victimizations in this study have not been narrowed to the sorts of things modeled in experimental simulations such as epileptic seizures, faints, falls, or smoky vents, but to the whole variety of circumstances captured when cameras are focused on the public in everyday life.

Putting events in perspective

When we reflect back to the incident that initiated this wave of ingenious experiments on bystander behavior, several issues seem to require more comment. Certainly, the attack on Catherine Genovese, her assault and stabbings clearly were dangerous emergencies. Contrary to the initial story, there were significant individual instances of helping behavior, as well as deliberate refusals to become involved. The story of the indifferent thirty-eight witnesses has been discredited. No one appears to know where the estimate of thirty-eight witnesses ever came from, although it seems that it originated with the police, and was fed to the Police Commissioner who planted the seed in the *New York Times*' editor A. M. Rosenthal, who commissioned the story that was researched and written by Martin Gansberg. Darley and Latané accepted the account at face value and reproduced in the laboratory a scenario that corresponded to the discredited *Times'* story. They created something in the lab that did not occur in everyday life. When we reflect on the way in which they constructed "groups," those individuals who were reluctant to intervene to help a stranger in distress, is it reasonable to presume that such entities coalesce in under six minutes? That was the laboratory framework employed to estimate the degree of mobilization of strangers. And the events which were created in retrospect were not so evidently dangerous to the observers. Their finding of "group" apathy may have been due to the failure to capture events in the laboratory that were as provocative as what occurred in real life, and which tested the resolve of persons enervated by such crises. The event was originally framed by the Police Commissioner as a story of urban apathy, with the underlying parable of urban dissensus and indifferent communities. But the most recent evidence from social psychology suggests that such a harsh verdict on interpersonal relations in the case of dangerous emergencies is not substantiated by contemporary research.

What if the Commissioner had framed the lessons of the murder differently? That is the question raised by Fran Cherry (1995). Cherry points to the way that gender, class, and historical conditions put lenses on how social psychologists approach their subject matter. She rejects the idea that psychological inquiries are undertaken in a perspective of impartiality and objectivity. The "stubborn particulars of social psychology" are those elements in our lives that surface whatever our inquiries assume consciously because they are so influential in shaping our experiences. Where Darley and Latané, following in the path laid out by Commissioner Murphy, see witness apathy, Cherry sees violence against women. Why does the social psychology of bystander indifference never reflect the fact that Winston Moseley attempted to rape Catherine Genovese? He stalked women at night in order to violate them sexually and murder them. He raped and murdered an African-American woman, Annie Mae Johnson, a month earlier in South Ozone Park and fifteen-year-old Barbara Kralik in Springfield Gardens the previous year (Kassin 2017). Nothing in the bystander apathy narrative points to the victimization of women. No one takes this scenario into the laboratory to model it, to parse its constituent elements, and to offer policies to abate this form of victimization (Baker 2014). Cherry's concern is what elements of the triggering event are imported into the narrative and become the subject matter for scientific examination and simulation – and what elements are disappeared from the original situation.

One of the things that was disappeared from the official narrative was Genovese's bisexuality. She was previously married to a man in a relationship that was annulled and, in 1964, was living with her female lover at the time of her murder. This never made it to the papers or to the social psychology laboratories. However, Cook (2014) reported that New York detectives questioned Genovese's partner, Mary Ann Zielonko, as a suspect in the murder. They believed that jealousy among gay people was especially intense and infidelity was a probable motive for violence. Zielonko was fast asleep at the time of the attack, and apparently a devoted partner. Karl Ross, a neighbor of Genovese and Zielonko who socialized occasionally with them, was also gay and was consequently reluctant to alert the police to the murder because it might expose him to police homophobic harassment. He witnessed the stabbing of Genovese in the vestibule.

Even if the salient issue in the Genovese event was the tension between individual versus collective responsibility, it is significant that the theoretical deep lifting went to the issue of the alleged failure of the bystanders to get involved. Why was the question not posed differently to highlight the situation of *the individuals*? Why do individuals respond? Is it because they all possess executive self-direction, that is, what is called self-control, self-management, and an ethic of responsibility? In an assembly of strangers, each one of whom has some level of self-control (i.e., an executive function), the initiation of collective action through leadership or social influence requires mastery of two things: cooperation of the other individuals and amelioration of the risk to the victim through a proposed course of action. Solo actors do not have to consult one another. What the experimentalists did not appreciate was the fact that solo action might be more expedient – not that groups are apathetic. However, in apprehending a perpetrator in everyday life in the rape simulation reported by Harari et al., an individual probably welcomes an assisting hand when acting alone might not achieve an optimum outcome. If that gives us a fair assessment of the situation, we can put the parable of the apathetic neighborhood and indifferent bystanders to rest.

What else does the Genovese case tell us?

Saul Kassin (2017), like Fran Cherry, did not believe that the most interesting issue in the Genovese story was witness apathy. Kassin points out that when Winston Moseley was first apprehended, it was for burglary of a home in the Corona Queens neighborhood undertaken in broad daylight. After he was seen removing a TV from a house and loading it into his Chevy Corvair, he was approached by a neighbor who asked him what he was doing. He said that he was helping the owners to move. This person called another neighbor to learn if anyone knew of such a move. When he learned that no such move was planned, that neighbor called the police to report the burglary, and the first neighbor popped the hood of the Corvair and disconnected the rotary cap that delivers the power to the spark plugs. When Moseley returned to his car, it would not start, and he casually walked away. He was arrested a short time afterwards. The man whose case went down in social psychology history because of the indifference of bystanders was subsequently arrested because the bystanders to the burglary took rather dramatic action to ensure he would be apprehended (2017:379).

Kassin argues that the second and more important phenomenon illustrated by the Genovese case was that of false confessions and false imprisonment. When he was interrogated by police, Moseley confessed to three murders – Genovese, Johnson, and Kralik. He was asked about two other murders – Emily Hoffert and Janice Wylie, the so-called "career girl murders." These were two professional women murdered in August 1963 in their Manhattan Upper East Side apartment. He indicated that he knew nothing about them. In the Genovese case, he reported that he had disposed of her wallet and car keys near his workplace, something that proved true and corroborated his personal connection to the crime. At the time of his arrest, another individual had been arrested for the Barbara Kralik murder, Alvin Mitchell, and had confessed. This was the situation that Rosenthal had wanted to ask the police commissioner about when the latter raised the story of the thirty-eight witnesses. According to Kassin, "police had interrogated Mitchell seven times for over 50 hours, culminating in an all-night session that lasted nearly 13 straight hours before he capitulated" (2017:375). Moseley had described the knife that he used to stab Barbara Kralik in her bedroom in the middle of the night at her parent's home. It was found discarded near the house. Police were suspicious of Moseley's confession to the Annie Mae Johnson murder because it seemed to differ from the coroner's account of the cause of death. Moseley said he had shot her twice in the stomach and four times in the back, then raped her before setting her house on fire. The coroner had concluded she died of puncture wounds from an ice pick or a similar sharp object. To discredit Moseley's confession for Kralik's murder – and thereby to preserve the prosecution of Mitchell – authorities flew to Johnson's home state of South Carolina where she was buried and had her remains exhumed. "To everyone's astonishment … the local coroner confirmed Moseley's account. Ms. Johnson was shot six times with a 0.22 caliber rifle – just as he had said. Four bullets detected in X-rays, were removed from her body" (2017:375–76). None of the police statements of either Mitchell or Moseley was tape recorded. Moseley was never indicted in the Johnson or Kralik cases. Mitchell was acquitted of the Kralik murder in 1964, but convicted on a re-trial a year later of first-degree manslaughter. He served twelve years and eight months before being released (Kassin 2017:377). At the first trial, Moseley appeared as a defense witness and described in vivid detail how he committed the murder. But the confession that police extorted from Mitchell appears ultimately to have sealed his fate.

Kassin also discovered another false confession related to the Moseley investigation – the Hoffert and Wylie murders mentioned earlier. George Whitmore, a nineteen-year-old African-American man, was arrested by police and signed an exquisitely detailed sixty-one page confession to both murders, and another unrelated homicide. This was after twenty-six hours of constant interrogation. He never read the confession that Brooklyn detectives had written for him, but immediately recanted it. He had a substantial alibi that put him in another location at the time of the murders. The actual perpetrator was subsequently apprehended and confessed. Whitmore was exonerated after serving three years and was awarded $500,000 for wrongful conviction (Kassin 2017:378). New York governor, Nelson Rockefeller, banned the death penalty in 1965 as a result of the Whitmore (and other) cases. Whitmore's case was cited in the famous *Miranda* decision in 1966 as the most conspicuous example of police abuse of custodial confessions.[5] Just as Cherry asks how the issue of violence against women could be missed, Kassin's question is: how did bystander apathy become the most salient element that psychologists seized upon to model in their laboratories, and how could they have missed the deep miscarriages of justice associated with the investigation of Winston Moseley (Editorial 2017)?

Conclusion

In 1977, writing from his prison cell, Winston Moseley wrote an op-ed that was published in the *New York Times*. It was titled "Today I'm a man who wants to be an asset." He noted, regarding the murder of Kitty Genovese: "The crime was tragic, but it did serve society, urging it as it did to come to the aid of its members in distress or danger." In other words, the 911 emergency number that has become universal in North America is the “service” his crime contributed to society. It was one of the unintended consequences of his depraved behavior. “The man who killed Kitty Genovese in Queens in 1964 is no more … Another vastly different individual has emerged, a Winston Moseley intent and determined to do constructive, not destructive things.” In 2016, Kitty Genovese’s younger brother, Bill, made a documentary on her murder – *The Witness* (Solomon 2016). Before the film was completed, he contacted Moseley with the intention of interviewing him about what had motivated him. Moseley declined the invitation but wrote back a long letter to the Genovese family, explaining that he had not personally murdered Kitty Genovese but had been involved as a driver in a mobster assassination that went awry. Moseley died in prison in 2016 at the age of eighty-one.

Notes

1 The exception was the size of the community in which the subject grew up. The size of the community was inversely related to helping behavior ($r = -0.26$, $p. < 0.05$). (Latané and Darley 1970:117).

2 Is titanium oxide safe in everyday life? “Based on the experimental evidence from animal inhalation studies TiO_2 nanoparticles are classified as ‘possible carcinogenic to humans’ by the International Agency for Research on Cancer and as occupational carcinogen by the National Institute for Occupational Safety and Health.” See Skocai, Filipic, Petkovic, and Novak (2011:227). While this may not have been of much concern to the Ss who appeared once for the experiment, the situation of research assistants who served in numerous experimental trials is another matter.

3 In their book, Latané and Darley add several further conditions. The fourth is that the responders have to determine the form of assistance to give. Fifth, they must decide how to implement their actions (Latané and Darley 1970:32). And they have to have sufficient skills to intervene (1970:36).

4 www.nytimes.com/2016/04/06/insider/1964-how-many-witnessed-the-murder-of-kitty-genovese.html

5 *Miranda v Arizona* 384 U.S. 436 (1966) was a decision of the US Supreme Court that required police to inform all persons arrested of a crime of the reason for their detention, of their right to consult a lawyer to help them prepare a defense, and the right to be protected from self-incrimination. The decision documented abuses of process in which accused persons were threatened in order to pressure them to confess, were confronted with false evidence to influence their confessions, and were promised inducements of mercy as a result of confessions. Without respect of the accused rights as spelled out in Miranda, courts could exclude any information obtained by police from persons held in custody.

6 Social psychology engineers wealth and intelligence

The Hawthorne and Pygmalion effects

Introduction: worker productivity and childhood IQ as expectation effects

In this chapter, we examine the controversies associated with two well-known investigations: the Hawthorne study and the Pygmalion, or the IQ expectation, study. Both were field experiments, one in industry, the other in education. Both purported to discover new, non-trivial information about human nature of tremendous relevance to society. Both had high impacts and apparently long-lasting implications for those who participated. The studies collected vital information over a long period of time, contrary to the usual short-term laboratory studies such as those examined in previous chapters and were heralded as landmark accomplishments and advances in knowledge. Both ultimately attracted close scrutiny, which suggested that the main effects were based on very small numbers of subjects, that both were open to sound, contrary interpretations, and that both enjoyed an appeal, like other studies in classical social psychology, which suggested that they conveyed powerful moral sentiments of more gravity than the evidence on which they were based. In addition, both were in the genre of expectation effects. The Hawthorne study suggested that worker output was limited less by such material factors as fatigue and remuneration than by the social relationships created by a progressive work environment. Pygmalion argued that the intellectual development of children was limited less by their innate biology than by the social expectations of their teachers. Many people continue to subscribe to such beliefs today because they contain a kernel of truth, but the idea advanced here is that the foundations for such beliefs appear to rest on something other than the science on which they were originally based. If these ideas were sound, then human beings could design societies in which industrial productivity and human intelligence would be boundless – and who would want to hope otherwise? That line of thinking makes the discipline of social psychology unresponsive to negative findings while fixating on ideas with tremendous moral appeal that are more responsive to “common-sense psychology” than to scientific psychology.

The Hawthorne effect

The Hawthorne studies were the single most important exploration of the human dimensions of industrial relations in the early 20th century. They were undertaken at Bell Telephone's Western Electric manufacturing plant in Chicago beginning in 1924 and continued through the early years of the depression until 1933. The Hawthorne plant manufactured a variety of electrical equipment and its growth reflected the burgeoning home telephone market that developed in the 1920s. It employed 22,000 workers in 1927 but this number grew to 40,000 by 1930 (down to 7,000 by 1932), reflecting the huge expansion (and contraction) of telephone services during the Roaring Twenties and the Great Depression.

Personnel managers with the company undertook a series of experiments to explore the effects of various conditions of work on worker morale and productivity, including changes in illumination, humidity, and work rests. In 1928, the company sought the input of several external experts, including Elton Mayo of the Harvard Business School, and Clair Turner, a professor of biology and public health at MIT, to help them interpret the results of their studies. One of the peculiarities of this investigation is that it is not clear who advanced the initial hypotheses in these studies, and what predictions were attached to the various changes in the conditions of work. Like some of the early classic studies in interpersonal influence reviewed earlier, much here appears to have been exploratory. The Hawthorne plant had created an Industrial Research Division. The research was certainly initiated internally at Hawthorne by management personnel, including Bill Dickson, Harold Wright, George Pennock, and Mark Putnam, but the subsequent findings are published in reports by people drawn into the project after its initiation, people whose intellectual stature dates to their interpretations of the Hawthorne studies. The classic sources are Elton Mayo's *The Human Problems of an Industrial Civilisation* (1933) and F. J. Roethlisberger and William J. Dickson's *Management and the Worker* (1939). Roethlisberger was a student of Mayo's at Harvard, and the Roethlisberger–Dickson account of the research is usually held as the authoritative one. It appeared a decade and a half after the start of the studies, and it was almost spiked by senior management at Hawthorne, who were alarmed by the claims that the management team in the bank-wiring shop was virtually incapable of affecting worker output, let alone determining what would be reasonable levels of productivity. The studies at Western Electric are memorable because of the discovery of the "Hawthorne effect." What that effect was, how it occurred, and how it came to embed itself so effectively in the consciousness of social psychologists, are not well understood. The term Hawthorne effect appears to have been first coined by Paul Lazarsfeld two years after the appearance of *Management and the Worker* (see Sobol 1959:52).[1]

The illusion of familiarity

It is noteworthy that in Mayo's preface to *Management and the Worker*, he alludes to the fact that there was some misunderstanding associated with the findings at Hawthorne. He says that his own Lowell lectures, North Whitehead's *The Industrial Worker* (1938), and Roethlisberger and Dickson's earlier business school account created

> an illusion of familiarity when the Hawthorne experiment is mentioned. But this is illusion: many of us have long been aware that there is no sufficiently general understanding of the course that the inquiry ran, of the many difficulties it encountered, and of the constant need to revise and renew the attack on the diverse problems presented.
>
> (Mayo 1939:xi)

Management and the Worker was going to set forth the full record and reverse the illusion by providing a full account of the development of the experiments.

There is little doubt that *Management and the Worker* struck a nerve among professional psychologists and personnel directors. Writing in the *Personnel Journal*, Charles Slocombe (1940), director of the Personnel Research Foundation, called it "the most outstanding study of industrial relations that has been published anywhere, anytime." Stuart Chase (1941), writing to a general audience in *Reader's Digest*, declared it: "the most exciting and important study of factory workers ever made. There is an idea here so big it leaves one gasping." However, what that idea was and why it left people gasping were not actually clear.

Today, we refer to the "Hawthorne effect" to denote a situation in which the introduction of experimental conditions designed to identify key aspects of behavior has the inadvertent consequence of changing the very behavior it is designed to identify. When persons realize that their behavior is being examined, this changes how they act, often resulting in their exhibiting socially desirable traits. Obviously, such changes are of interest to psychologists who are trying to understand the rationale of behavior as it transpires in context, and who need to separate aspects of behavior that are natural or spontaneous from that which results from the conditions of experimentation.

The original Hawthorne effect referred to the *claim* – for, as we shall see, much here remains illusory – that the productivity of the workers increased over time with whatever variation in the work conditions was introduced by the experimenters. Where Heisenberg had noted in physics that the act of observation changed the field of observation, the Hawthorne effect suggested that this change was motivated by social considerations that led those exposed under the experimental microscope to put their best foot forward – to excel, to show themselves in the most positive light, to produce more, and weather the tribulations of industrial work with personal grace and dignity. The changes in productivity or output were a function of tacit expectation effects. In contemporary social psychology, this is referred to as "priming" (Molden 2014). The evidence for this was suggested in the preliminary illumination experiments and in the relay assembly test room.

The illumination and first relay assembly room tests

The illumination experiments were initially designed to determine whether increases in artificial lighting on the factory floor could result in fewer accidents, less eye-strain, and higher productivity. The electrical industry had a considerable investment in establishing the returns to enhanced lighting, and the U.S. National Research Council became involved with a blue ribbon panel of experts headed by Thomas Edison to explore the effects of changes in illumination. The experiments were conducted at the Hawthorne plant over a three-year period (1924–1927) and involved workers manually winding induction coils for telephone systems. It was clear to the engineers that identification of the contribution of illumination to productivity net of the effects of other changes created by the experimental conditions would be difficult. Baselines of productivity were taken, the women recruited were interviewed about the experiment, changes were made in illumination, and measurements in output during the day were taken by foremen to identify levels of productivity, along with measures of temperature and humidity A control group of workers not subject to the same enhanced conditions of supervision experienced increases in output because of the development of informal competition between the workers. Roethlisberger and Dickson provided a summary report of the study in the Introduction to *Management and the Worker.* They noted that even when light values were decreased, output increased. In fact, in one variation, even when the light was cut down to 0.06 of a foot-candle, "an amount of light approximately equal to that on an ordinary moonlight night … the girls maintained their efficiency" (1939:17). It appeared as though the physical conditions of illumination were less consequential than the psychological conditions. In the eleven periods of the third round of experiments, the control group as well as the experimental group showed an improvement from a prior baseline. This occurred whether the illumination was increased, decreased, or remained constant. In the end, "the results of these experiments on illumination fell short of the expectations of the company in the sense that they failed to answer the specific question of the relation between illumination and efficiency" (1939:18). But they did motivate the next phase of inquiry: the relay assembly room tests.

The relay assembly room tests started in April 1927 and continued until June 1932, when the demand for relays was so low due to the Depression that the study was terminated. The Roethlisberger–Dickson report covers the first thirteen periods, ending coverage in June 1929. This was the most famous part of the Hawthorne study, and the one that has received the greatest empirical scrutiny. It reflects the theoretical ideas derived from Elton Mayo, who suggested that, in industrial conditions, people are not motivated by simple physical conditions such as exhaustion or fatigue, and neither is their productivity determined primarily by their economic self-interest and material aspirations. Although fatigue and exhaustion were a concern in nineteenth-century conditions of production, heavy labor was increasingly replaced by machines and fatigue in modern workers was viewed as an expression of morale and workplace adjustment. As for income, Mayo suggested that beyond a certain level of material comfort, workers put more stock in the social dimensions of work. They valued the relationships between each other and their supervisors. Paramount was the importance of developing a humane set of relationships that recognized the total situation of the worker and her sentiments. Mayo drew heavily from Pareto and Freud, authorities who held that most behavior was not rational as classical economics had held. This point is important since most of the postwar criticisms of the Hawthorne study stress the role of self-interest in the level of productivity and in the restriction of productivity by pieceworkers out of self-interest.

The design of the relay assembly test room was more of a "test–change–retest" design based on the performance of a fixed number of workers whose output was examined under successively altered conditions of work. The relay assembly test room included a change of work location in which five operators and one layout worker assembled complicated relay switches consisting of thirty to forty parts in a separate test room. Other changes included a collective form of remuneration calculated on the combined productivity of the test room workers (piecework), introduction of break periods of various lengths at various times during the day, provision of lunch and beverages by the company, and alterations in the weekly work schedule (shortened days, shorter work week). The group was monitored by an observer who came to act in a cooperative supervisory capacity with the workers. The output was calculated mechanically with a ticker tape machine, and by manual summaries. The observer made notes about the small talk and social interaction of the workers. To induce them to attend hospital for regular medical assessments, the workers were bribed with ice cream. The workers brought their own cake – and as the experiment developed, the workers, all young women in their late teens and early twenties (with one exception), began to socialize outside the workplace.

The analysis of the changes in productivity is quite detailed but the conclusions were quite simple. If one examines the average hourly output per week during the first thirteen periods of the study, including a reversion to the standard regime during period twelve (in which pauses were cancelled and the work week lengthened), the level of output drifts haltingly upward period after period. "Examination of this chart reveals at once no simple correlations between the experimentally exposed changes in working conditions and rate of work" (1939:75). From a baseline of around fifty relays an hour in the first weeks, the women increased their output to sixty or seventy relays per hour two years later. The experimenters noted that the workers appeared to have become a spot "healthier" as gauged by a gain in weight over the period. They also took pains to rule out a decline in fatigue as the cause.

The second relay assembly and the mica-splitting tests

To tackle the potential contribution of changes in the wage incentive, they created a second relay assembly test group and a mica-splitting test group. The former worked together on the normal shop floor with the normal form of supervision. For a period of nine weeks, they received the small-group piece rate, then reverted to the shop-wide form of remuneration. Two operators continued to report inflated productivity after a return to the old method of payment and two did not – from which Roethlisberger and Dickson inferred that "it was difficult to conclude whether the increase in output was an immediate response to the change in wage incentive" (1939:132–3). However, they also reported that, because of friction on the shop floor between the special group and the rest of the workers, the foreman demanded that the former method of payment be reinstituted. So, this manipulation was inconclusive.

The mica-splitting test group was similar to the original relay assembly group except for the change in method of remuneration: in other words, they were isolated in a test room, but earned the general piece rate of the other workers. This test began in 1928 and terminated in mid-1930 when the demand plummeted and, with it, productivity declined for want of work. For this reason, Roethlisberger and Dickson employed only the first fourteen months of the two-year series (1939:153). Even with this truncation of the series, the gains in productivity were modest and were inconsistent across the different workers.

> In both test rooms, output tended to increase in the first year. Also, in both cases the increases followed experimentally induced changes in work conditions. With these two exceptions, however, no parallel developments in the two rooms could be detected
>
> (1939:149)

And again, these changes were estimated on the bases of a handful of workers – five in the case of the relay assembly and five in the mica-splitting test room.

Search for the real Hawthorne effect

With these inconclusive observations, Roethlisberger and Dickson went into completely different methodological directions, which led to a deeper understanding of the Hawthorne effect than that with which they started. They focused on the qualitative data from the second relay and mica test groups, which pointed to the dramatic difference in the social situation between the initial relay assembly test group and the subsequent two groups. The original relay assembly group had developed a rare industrial tone in which workers did not feel harassed by their bosses. Indeed, they did not view the observer as a boss at all. The atmosphere was one of a new employee–supervisor relationship marked by a spirit of cooperation, in which "there were no longer any bosses." Absenteeism declined. Group morale improved. Everyone was more likely to assist the others. By contrast, in the later studies with their more modest improvements, there was an "apprehension of management," and a fear of unemployment as the "dreaded depression" (1939:153) began to make the future uncertain. That magic first glimpsed in the illumination studies and corroborated in the relay assembly vanished. Having established to their own satisfaction that productivity was not powerfully linked to wages, Roethlisberger and Dickson began to outline the real Hawthorne effect.

The real Hawthorne effect was the potential change in industrial relationships made possible by the insights of scientific management of the sort proposed by Elton Mayo. The bulk of the 600 pages in this classic book is not devoted to the relay assembly test and its seemingly irrepressible increases in productivity. Instead, we find a program based on widespread interviewing, understanding the nature of industrial conflict based on a novel theory of human nature, and devising a profession capable of achieving industrial harmony through reliance on a scientific management approach that bordered on psychiatric therapy. The interview phase involved some 21,000 employees. It followed an incredible logic of expansion as the short two to three pages of notes gathered in twenty-five-minute interviews exploded into dozens of pages of transcripts. Interviews increasingly came to resemble therapeutic sessions lasting for hours, as the interviewers were alerted to the latent content of worker cognitions in search for the "total situation" of the workers. That material has never been analysed.

Tests in the bank-wiring room

The bank-wiring room was devoted to the creation of large electrical switching appliances. The labor force was male, and worker output consisted of two units per day. This was referred to as "the bogey," an informal level of productivity enforced by the workers through informal social control, verbal taunts, and playful shoulder punches. There was no new management intervention and nothing of the magical change in productivity associated with the relay assembly test room. The inclusion of this analysis in *Management and the Worker* is difficult to understand, since the conclusions here have none of the implications of the relay assembly room study.

Criticisms of Hawthorne

The conflict over what I would call the small "h" Hawthorne effect arises in a number of works. One of the most provocative is owed to Alex Carey:

> A detailed comparison between the Hawthorne conclusions and the Hawthorne evidence shows these conclusions to be almost wholly unsupported. The evidence reported by the Hawthorne investigators is found to be consistent with the view that the material, and especially financial, reward is the principal influence on work morale and behavior. Questions are raised about how it was possible for studies so nearly devoid of scientific merit, and conclusions so little supported by the evidence, to gain so influential and respected a place within scientific disciplines and to hold this place for so long.
>
> (1967:403)

Carey's point was that the small-group wage system was the primary cause of worker output among the five subjects in the relay assembly room. The more units they assembled, the higher their pay. Humanitarian management was beside the point, and the so-called Hawthorne effect was a myth.

Carey and other critics pointed to the replacement of two operators in the relay assembly test room who were repeatedly criticized for "excessive" talking and who seemed to be consciously limiting their output in contrast to their peers. Despite the fact that they had been led to believe that enhanced productivity was not an objective of the experiment, they were returned to the shop floor for a lack of cooperation on the issue of output. They were replaced by two new operators whose productivity immediately led the pack: one worker, a fifteen-year-old Italian woman, was particularly productive, and apparently cajoled the others to monitor and elevate their output. As Carey and others have pointed out, her mother had died, and her brother and father were facing unemployment. The observer's record provides fairly convincing evidence that she was instrumental in trying to elevate levels of collective output – a fact from which everyone in the test room benefited materially. The importance of this instrumental action was neutralized by the move to recover the whole technical and human situation associated with the interview phase. This led Carey to ask:

> How is it that nearly all authors of textbooks who have drawn material from the Hawthorne studies have failed to recognize the vast discrepancy between evidence and conclusions in these studies, [and] have frequently misdescribed the actual observations in a way that brings the evidence into line with the conclusions?

Over two and a half decades after Carey's critique, Stephen Jones again examined the evidence for a Hawthorne effect, that is, a change in productivity unrelated to economic factors, in an article entitled "Was There a Hawthorne Effect?" Jones modeled week-to-week output by regressing the average hourly rate of productivity on both formal changes introduced by management (form of remuneration, rest breaks, etc.) and inadvertent changes brought about by circumstances (loss of productivity due to repair time, unemployment, radical changes in temperature, etc.). He tested for immediate effects and delayed effects occurring up to four weeks later. Again examining the relay assembly room test data, he concluded that

> contrary to the conventional wisdom in much research and teaching, I have found essentially no evidence of Hawthorne effects, either unconditionally or with allowances for direct effects of the experimental variables themselves. My results appear to be robust across a wide variety of specifications, alternative samples, and two definitions of experimental change.
>
> (1992:457)

H. M. Parsons (1974), writing in *Science*, had attributed the changes in output to operant conditioning – the fact that the workers could constantly monitor their output and benefit from changes in collective productivity, resulting in a long-term increase in skill levels. Parsons and Jones differ in emphasis because the model Jones tests incorporates a simple stepwise change in the method of payment (the mean for change was 0.97, i.e., unity, or no variability).[2]

Jones ends his observations with the following: "A fruitful line of sociological inquiry … would explore the social and historical context whereby the Hawthorne effect has become enshrined as received wisdom in the social sciences" (1992:457). In fact, that is what Richard Gillespie tackled in his book, *Manufacturing Knowledge: A History of the Hawthorne Experiments* (1991). He takes the position that a conspiracy of Harvard management professors and industry managers emerged to marginalize the relevance of the economic aspirations of workers and their grasp of their working environment.

If my instincts are correct, the reason that Hawthorne persists in the imagination is because it paints a picture of workers and industrial production of heroic proportions. The workers are cast as subjects prone to morbid fantasies that they are hardly capable of understanding. Their complaints to management have to be interpreted in terms of the total situation both on and off the job, in terms of both manifest and latent content. Complaints against management that were discovered in the interview phase often reflected the obsessive thinking of workers – and the analyst risked superficial reduction of feelings of personal insecurity and morbidity to elements of the workplace thought by Mayo to be incapable of explaining them. The scientific manager has to coordinate the technical and the human facts of production to maintain both a personal and social equilibrium. So, the task for the industrial psychologists was not simply about improving the manufacture of widgets, that is, production and efficiency (i.e., the empirical illusion of Hawthorne), but maintaining social integrity in an industrial system prone to destroying it, or prone to undermining productivity by failing to account for the human factor (as in the bank-wiring study). Understood in this way, the Hawthorne effect was not a methodological artifact, as we have come to view it consequently. It was the clue to social transformation through expert psychological knowledge glimpsed by a mentally healthy work force operating at optimum levels of achievement in the illumination and relay assembly room study. It was about changing civilization by integrating the technical engineering in the manufacturing process (efficiency, productivity, workmanship) while steering workers wide of their obsessions and morbid thinking. That is what made the idea so big it left people gasping. However, at the same time, the dismissal of the economic consequences of the small-group wage system by the authors left the critics shaking their heads. Here may be the key to the persistence of Hawthorne. The moral attraction of the idea finds its continuing relevance as a vision of the humanization of industrial civilization – the quest for paradise in the age of fragmentation – while the empirical evidence points perennially to its negation. (Paradise lost?) Professors of management and industrial psychologists appear fixated on the cultural ideal of Hawthorne's industrial magic and return to the oracle of the relay assembly room test with fascination – only to find the divinations ambiguous. As a result, students of Hawthorne have accorded it a place of pride in the arsenal of industrial psychology, not because of its accomplishment, but because of its promise. Mayo's observations in the preface to *Management* support this line of inference:

> The art of human collaboration seems to have disappeared during two centuries of quite remarkable human progress. The various nations seem to have lost all capacity for international cooperation in the necessary tasks of civilization. The internal condition of each nation is not materially better. … In this general situation it would seem that inquiries such as those undertaken by officers of the Western Electric Company have an urgent practical importance that is second to no other human undertaking. How can humanity's capacity for spontaneous cooperation be restored?
>
> (Mayo 1939:xiv)

Here was work that was practical, timely, and inspired. And little supported by the evidence. In retrospect, it appears that industrial studies were incapable of dispersing the dark clouds of history settling over Europe when *Management* was published. But it might have represented *an analogy* to Mayo as to how the conflicts in what would become the most fearsome war in living memory might be defused. A relevant lesson could be drawn, however remote from the primary study. The study promised a brighter future. The heroic achievements of five young women in the relay assembly room provided a template for prosperity (as well as peace). In today's schools of business, the myth of Hawthorne survives the deficiencies of evidence noted by Carey, Parsons, Jones, and others.

In my view, there are strong parallels for the improvement of society with the later investigations of intelligence in the Pygmalion research. Where Hawthorne might unleash unprecedented levels of industrial productivity by creating an optimal working environment and a harmonious industrial community, IQ could be cultivated in an educational environment designed to make it grow unencumbered by extraneous obstacles and prejudices. In this way, racial and ethnic minorities could move forward. This was the kernel of thought behind Robert Rosenthal's famous study, *Pygmalion in the Classroom* (Rosenthal and Jacobson 1968). Where Mayo alluded to a solution to international conflict based on a template of industrial ideals, Pygmalion promised an end to racial and ethnic inequalities by addressing the impediments to social advancement created in educational institutions.

The Pygmalion effect

In the Greek myth, Pygmalion was the king of Cyprus who, it was said, created a beautiful female figure in ivory – Galatea – for whom he pined until the figure was brought to life for him by the goddess Aphrodite. This story supports the notion of wish fulfillment where human desire can make the improbable happen. George Bernard Shaw's *Pygmalion* is the story of East Londoner Eliza Doolittle, whose cockney accent marks her lower-class origins, resigning her to a fate of poverty, until her fortune is reversed by language training under the tutelage of Shaw's eccentric Dr. Higgins.

En route to teacher expectancies and IQ studies, Rosenthal examined a family of behavioral studies that pointed in a similar self-fulfilling direction where social expectations brought about the situation they initially only imagined. In the case of Clever Hans, the horse owned by Wilhelm Von Osten, a German mathematics teacher, Pfungst ([1911] 1965) traced the animal's remarkable abilities to the tacit communication of expectations by those who put various questions to the animal. Visitors asked the horse questions of addition, subtraction, multiplication, and division (the answers to which they knew) and the horse answered by tapping his foot to the appropriate number. The tracking of the animal's movements by the audience gave the horse clues about when to start and stop the exercise, creating the impression of great ability. Hans followed the questioners' expectations, although the latter were unaware of their own tacit signals to the animal.

Rosenthal reports that students given what they believed were "maze bright rats" were able to teach the animals maze discrimination tasks quicker than students given what they believed were "dull" rats. But the rats were not in any way different. This led Rosenthal to suggest the foundation for the subsequent experiment with children: "If rats become brighter when expected to then it should not be farfetched to think that children could become brighter when expected to by their teachers" (Rosenthal 1985:44). *Pygmalion in the Classroom* provided evidence that this occurred although, like Hawthorne, such a strong claim attracted close scrutiny and a growing body of skepticism.

But surely Rosenthal's line of thinking was illogical from the start. Did Clever Hans become literate because people treated him so? Certainly not. Did rats get smarter? Just as subtle clues in the case of Clever Hans suggested to the animal how to behave, why should we not permit the same explanation in the case of the maze bright rat study? Students with so-called "bright" rats could simply have taken more time to permit them to do their maze runs. They may have handled them more defensively, and, indeed, they must have treated them differently because "smart" and "dull" rats were all actually interchangeable, according to Rosenthal. In addition, their conduct may have actually been observed and scored differently during their performance because of the different expectations. But none of this means they were actually "smarter" rats after five days in the maze in the hands of these inexperienced students.

How can we infer that differences in performance mean real differences in learning when differential expectations are confounded by different handling? And even if one were prepared to go out on that particular limb, why should we equate differences in learning under these circumstances with differences in "brightness"? The entire foundation for the research is erected on sand (Jarrett 2008).

As with Hawthorne, the underlying theory is murky. It conflates changes in the experimenter's cues with changes in the subject's capacity, even though the prior case studies – except at the level of myth or analogy – fall decidedly short of that.

Evidence of the impact of teacher expectations on pupil IQs was first announced at the end of Rosenthal's book, *Experimenter Effects in Behavioral Research* (1966).[3] It was reported more fully in *Pygmalion in the Classroom* by Robert Rosenthal and Lenore Jacobsen (1968). However, Rosenthal was also involved in four other tests of the hypothesis published between 1968 and 1974. Like Hawthorne, this work was extremely relevant to everyday life. It provided a novel explanation of differential patterns of school success by questioning the role of teachers in cultivating the basic raw talent of students under their charge. It was a long-term field experiment, covering a period of about two years, with superior promises of ecological validity. Because of the nature of the dependent variable and the length of the study, it promised to be a high-impact study with significant consequences for the subjects. And the specific hypothesis about how teacher expectations might influence student IQ certainly was not a common-sensical hypothesis in search of anecdotal support.

The Pygmalion study appeared in a highly charged ideological context in which sizable investments of public money were being poured into "headstart" and remedial education programs to alleviate the dramatic levels of school failure among poor people, particularly poor black and Hispanic communities, and to reverse the cycle of poverty and racial alienation in modern society. Rosenthal's perspective put a new interpretation on the relationship between poverty and school failure. Rather than arising from a lack of home resources, a lack of parental support, or a lack of home schooling prior to public schooling, the theory of expectation effects shifted the blame for school failure to teachers. Poverty might be related to school failure because middle-class teachers (both black and white) tacitly prepared poor students (usually minority group members) for failure because they *expected* them to fail. The expectation might work as a self-fulfilling prophecy. The political implications of the research were entirely unanticipated by the previous line of inquiry and attracted uncommon levels of public attention.

The study: manipulating positive expectations

To explore the self-fulfilling prophecy, Rosenthal and Jacobsen manipulated teacher expectations in the Spruce School of South San Francisco. The school covered kindergarten to grade six. Each class level was divided into a, b, and c levels, reflecting above average, average, and below average performance levels. Students completed a little-used IQ test – Flanagan's Test of General Ability. This examined both verbal and reasoning abilities and was differentiated for various age groups (K-l, 2–3, 4–6). Its introduction into the schools was disguised as a "Test of Inflected Acquisition." Supposedly based on a joint Harvard–National Science Foundation study, the teachers were told the following:

> As a part of our study we are further validating a test which predicts the likelihood that a child will show an inflection point or "spurt" within the near future. This test which will be administered in your school will allow us to predict which youngsters are most likely to show an academic spurt. The top 20% (approximately) of the scorers on this test will probably be found at various levels of academic functioning.
>
> (Rosenthal and Jacobson 1968:66)

The test scores purportedly permitted the examiners to predict spurts in IQ gains in certain students. Jacobsen was principal at the Spruce School and none of the students, their parents, or teachers were advised that they were part of an experiment. Hence, no one consented to be studied in this fashion. Approximately 20% of the students were identified to teachers in the start of the fall term several months after the initial test. Rosenthal and Jacobsen report that the "bloomers" were chosen at random.[4] The test was administered repeatedly to measure changes in IQ. In fact, the test was administered four separate times: (1) in May 1964 to establish a baseline, (2) at the end of the fall term to establish any immediate effects, (3) at the end of a first year to establish the basic post-test results, and (4) at the end of the second year to establish the long-term post-test results.

The results fall into three areas: aggregate changes in the IQ of pupils reported for the experimental and control pupils by class, changes in the school grades by subject, and changes in teacher attitudes to the students. In terms of IQ changes, the following was reported. First, after one school term, there was some evidence that the experimental group as a whole showed an IQ increase of 2.29 points, but this was not statistically significant ($a = 0.08$).[5] Second, after a full year, there was an overall IQ gain of 12.22 points. However, the effects were based on the performance of seven grade one and twelve grade two students (formerly K and 1 in the pre-test).[6] Examining these nineteen students, 79% experienced at least a ten-point gain, 47% experienced a twenty-point gain, and 21% experienced a thirty-point gain.[7] None of the other grades showed any significant differences. Third, as for the long-term effect, after two years, the expectancy advantage was non-significant for the younger students, but the students in grade five showed evidence of dramatic gains – 11.1 points.[8]

In terms of academic subjects, there was evidence of an expectation gain for the experimental subjects in the first three grades after one year. However, it was found for only a single subject – reading – and the scale used to present the differences was calculated in tenths of a letter grade, effectively magnifying small differences.[9]

The final area of measurement concerned attitudes. Rosenthal and Jacobsen compared teacher attitudes toward their experimental and control subjects.[10] The experimental subjects were thought to be more curious, more interesting, more likely to succeed, more appealing, better adjusted, happier, etc. In fact, Rosenthal and Jacobsen reported that where control subjects experienced significant IQ gains, there was some evidence of an attitudinal backlash from the teachers – students were given lower evaluations on these dimensions where IQ was not expected to improve. It is difficult to gauge the theoretical relevance of this part of the work. Certainly, Rosenthal had earlier asked the student experimenters in the rat study about their attitudes toward rats – presumably to establish that at some level the expectation effect (dull versus bright) had resulted in differences in orientation toward the rats. What is so difficult to gauge here is why evidence that appears so strongly to suggest differential attitudes would be found among teachers who, after a year, could hardly remember which pupils were supposed to be the "bloomers." And if the attitudinal shift was so vivid, it is surprising that it impacted only a single academic subject, and only when the grade range was stretched beyond credulity.

The explanation

There were no actual observations made of how teachers treated the subjects. Rosenthal and Jacobsen had to speculate about the mechanism by which the expectations were actually transmitted. Was it the case that the teachers spent more time with each of the students whose IQs were expected to spurt? Probably not – since where the experimental subjects' IQs jumped, so did that of the class as a whole – suggesting that the teachers were not investing time exclusively with the "bloomers." Were the teachers *talking* more to "bloomers"? And was this the way the expectations were transmitted? When we look at the evidence, we see that the greater IQ gains were made in reasoning IQ for both groups.[11] In one year, the experimental group jumped 22.86 points compared to the 15.73 points for the controls. So, this line of thinking, according to Rosenthal and Jacobsen, seemed improbable.

According to Rosenthal and Jacobsen, higher expectations must have been transmitted by tone of voice, facial expression, touch, and posture. How this worked is a matter of speculation. These tacit expectations may have impacted the self-concept of the subjects, yielding better performance outputs, higher practice and exercise of abilities, and, ultimately, better performance on the IQ test. There remained two further problems for Rosenthal and Jacobsen. First, why was the basic effect found only for the *youngest* pupils and, two, why was the long-term effect found only for the *older* students? Rosenthal and Jacobsen argued that the young students were more plastic, more malleable, and easier to influence, but also required ongoing reinforcement to sustain the change. As for the older students, if the message did get through, and somehow escaped measure initially, it might survive longer since the older students, because they were more set in their tendencies, would not require ongoing reinforcement. All this was possible, even if it was *ad hoc*.

Impact of Pygmalion in the classroom

The Pygmalion study received first prize for the Cattell Fund Award for experimental design given by Division 13 of the American Psychological Association in 1969 and the book was reviewed nearly universally in the contemporary press. Rosenthal and Jacobsen (1969:23) summarized their findings in *Scientific American* and pointed to the need to focus attention on the way teachers structure the performance of students. The study was discussed in the *New York Times*, *New York Review of Books*, *Times Literary Supplement*, *Saturday Review*, *New Yorker*, and many other popular periodicals. Rosenthal was interviewed by Barbara Walters on NBC and the book quickly became standard reading at colleges of education throughout North America and Europe. At the time, millions of dollars of federal money were being poured into ghetto education for headstart programs, remedial programs, and cultural enrichment. At the time, there was evidence that such spending *did* boost IQ performance of disadvantaged kids – one study cited a ten-point gain for 38% of students, and a twenty-point gain for 12% of students – but this was over three years. Compare this to the ten-point gain by 79%, twenty-point gain by 47%, and thirty-point gain by 21% in one year in the Pygmalion study! These were spectacular increases and they lent credibility to the notion that racial and class differences in educational accomplishment could be explained in large part by how teachers treated their students. This was the heyday of labeling theory. Given all the attention that it received in the popular culture and in the academy, it was not long before critics ground down their microscopes to examine Pygmalion more closely.

The critical responses

Rosenthal mentions casually that

> the bulk of the negative reactions [to Pygmalion] came from workers in the field of educational psychology. Perhaps it is only they who would have been interested enough to respond. But that seems unlikely … We leave the observation as just a curiosity
>
> (1985:49)

to be clarified by historians, sociologists, and psychologists of science. In other words, opposition appeared for apparently extra-scientific reasons. Readers can judge for themselves whether science was not better served by these skeptics than by all the yea-sayers who did not want to look too critically at the evidence. The comments reviewed here derive from several now-classic critiques of Pygmalion, including Thorndike (1968), Elashoff and Snow (1971), and Cronbach (1975). In his review of the book in 1968, Robert Thorndike wrote:

> In spite of anything I can say, I am sure [Pygmalion] will become a classic – widely referred to and rarely examined critically. Alas, it is so defective technically that one can only regret that it ever got beyond the eyes of the original researchers.
>
> (reprinted in 1971:65)

Elashoff and Snow reported in a similar vein that, despite the attention the book received in official circles, “We retain our view that *Pygmalion* was inadequately and prematurely reported to the general public” (1971:161). Wineburg recorded that

> even before the book hit the streets, headlines about it splashed over the front page of the *New York Times.* … Details of the experiment's failure to replicate, however, received a scant column inch in the continuation of the story on page 2.
>
> (1987a:31)

Recall why the study was attractive to begin with. Producing changes in performance of a short-term nature, as in hypnosis, or compliance to bizarre short-term demands in an artificial laboratory setting, is one thing. But IQ is not plastic and it is not voluntaristic behavior. It appears to be more or less fixed. Control of something like IQ by the act of volition would be impressive if, in fact, that is what occurred.

The problems

There were major problems having to do with the way in which the Test of General Ability (TOGA) was administered. Recall that Flanagan's TOGA, disguised as the test of "Inflected Acquisition," was used partly because teachers might have been more familiar with the Stanford–Binet IQ test – which was the sort of instrument in use in professional educational circles to diagnose learning problems. The TOGA test was created for three grade levels: K-1, 2–3 and 4–6. This means that, over the course of the experiment, the same base intelligence would be estimated by three different instruments as the children got older. Imagine the tests administered in spring term over three years. In reporting this, I borrow from Thorndike's review. Note that as the children are tested at different dates, they are examined with different versions of TOGA (Table 6.1).

Because there are different versions of the test, different outcomes may be a result of different test measures, not changes in IQ. This is a problem of

Table 6.1 Changes in version of test as students changed grades

*Grade (initial test)**	*Version of test between grades*	*Grade (second test)***	*Version of test between grades*	*Grade (third test)****
K	Same test	1	Different test	2
1	Different test	2	Same test	3
2	Same test	3	Different test	4
3	Different test	4	Same test	5
4	Same test	5	Same test	6
5	Same test	6		Junior High School

* Spring of first year (pre-test)

** End of the first year (basic test)

*** End of the second year (long-term test) "reliability." Usually, when tests are administered to large numbers of people, it is possible to identify the degree to which various versions of the test measure the same ability. The "concordance" between different versions of TOGA were not known because the test was not in wide use (and was attractive for that reason in the experiment). Thus, the "consistency" in what was measured was open to question. In addition, if one looks at the older groups, they tended to take the same test repeatedly. This suggests an effect arising from practice. One group would have taken the identical test four times and this group would show the greatest long-term gains in IQ. Obviously, this is alarming. However, Rosenthal replied that while these deficiencies may have been real, the important thing to look at is the *differences* measured between control and experimental groups – and these were significant even if allowing for practice and inconsistency in the measures.

Thorndike points out that none of the published reports of the study contained the original TOGA scores. The appendices to the report contained the average pre-test scores by class for reasoning and verbal IQs. The text reports the "difference scores" calculated by subtracting the post-test means from the pre-test means. The case for prophecy effects is based on the performance of the first two grades – specifically, the performance of seven experimental subjects in grade one and twelve experimental subjects in grade two. In Rosenthal's appendix tables, one finds classes with average IQs of 31, 47, 53, and 54. An average IQ is 100. Thorndike notes that these classes just barely appear to make the grade as imbeciles. And yet these defective pre-test data were used by the authors without caution as to their validity. Thorndike recalculated the average verbal and reasoning IQs combining all three levels (a, b, c) for the first and second grade. They are shown in Table 6.2

What kind of test is it that gives a mean reasoning "IQ" of 58 for the total entering a first-grade class in an ordinary school? The pre-test data are highly suspicious. If IQ = mental age/physical age and if IQ = 0.58, then it is possible to estimate the mental age of the children. If we assume that on the pre-test physical age = 6, then mental age = *x/6* = 0.58. Solving for *x,* we deduce a mental age of 3.5. What score on the original TOGA was required to achieve a mental age of 3.5? The tables do not report for such low ages. Estimating a mental age of 5.3, one would need to score 8 out of 28 items. Again, extrapolating downward, "we come out with a raw score of approximately 2! Random marking would give 5 or 6 right!" (Thorndike 1968:710).

Table 6.2 Average verbal and reasoning IQs for first and second grade students

	First grade	*Second grade*
Verbal IQ	105.7	99.4
Reasoning IQ	58	89.1

It is reckless to base any inferences on a test that is clearly so suspect in the identification of its baseline. That is something that most readers would never recognize because the report describes "difference" scores. The raw scores are nowhere produced and initial and post-test scores were put in appendices. Rosenthal (1969) explained that the low pre-test scores were not inaccurate. "These low IQs were earned because very few items were attempted by many of the children" (1969:690). When one examines the items in question, it is hardly surprising. The reasoning IQ questions displayed abstract geometrical forms with the instruction: "find the exception." Given that the children would hardly be able to read, let alone distinguish asymmetrical line puzzles, the pre-test takes on a different significance. Basically, the children were barely literate on the pre-test reasoning questions and did much better a year later when the test made more sense to them. On this Wineburg notes:

> A change on the pre-test may be interpreted as "intellectual growth," but given what we know about the pre-test, we could just as easily attribute it to other factors – misunderstood test instructions, uncontrolled test administration, selective teacher coaching, teacher encouragement for guessing, or even chance.
>
> (1987b:43)

Pesky educational psychologists

Thorndike tackles similar anomalies in the post-test scores. Table A-6 reports that, for one classroom, there are six pupils with an average IQ of 150 points and a standard deviation of 40 points. Again, we can estimate the mental age. At the end of grade one, if we assume that the children are 7.5 years old and if IQ = mental age/physical age, then the mental age of students with an IQ of 150 is 11.25. Again, we ask what do the scores of students with a mental age of 11.25 look like? The tables only go to a mental age of ten – and at that level the students would score 26 out of 28. Students with a mental age of 11.25 have to score even higher – but they are already approaching perfection (more than 26 out of 28). With such scores, what is the meaning of a standard deviation of 40 points? The data are so untrustworthy as to make inferences based on them reckless. As Thorndike advised, when the clock strikes thirteen, pitch it!

Many of these problems apparently escaped the readers of the original work. First, the report is based on difference scores, not the initial raw scores. Second, the figures that showed the dramatic spurts in IQ (79%, 47%, and 21%) did not always indicate the small sample sizes on which they were based. And, finally, shifts in academic performance, namely reading, were represented as microscopic differences that reported one-tenth grade scales that overemphasized minute differences.

Rosenthal's response

Rosenthal replied to Thorndike's criticisms (and others) by arguing that if the pre-test scores were unreliable, this made the measurement of differences between control and experiment groups harder to obtain, but the tests find such differences to be (sometimes) significant, and in the directions predicted. In other words, even if one allows that there are issues of validity, this does not entitle one to dismiss the differences that were accurately predicted between the groups. However, this is debatable. The major gains come from the first two classes and are based overwhelmingly on the reasoning component. These are conditions that contributed uniquely to the measured gains (i.e., they were not found in these classes in the reading IQ dimension, or anywhere else in either dimension). Aside from differences between the groups arising from the unobserved expectation effect, we know that a large component of the change arises from a comparison of performance before and after the children learned how to read, fill in the blanks, take tests, and meet other academic expectations. We also know that in all the other classes where the pre-test means were normal, there was no expectation effect. This puts us in the position of attributing all the difference in reasoning IQ in the first post-test to the expectation effect – which assumes we can sensibly subtract a valid score from the post-test measures from the completely meaningless pre-test score. In my view, that is foolhardy.

Rosenthal's advice is also perverse in view of his own failures to replicate the same test. We are asked to take the evidence from one study in five where the gains are discovered for a minority of classes and identified inconsistently over two periods of measurement. Evans and Rosenthal (1969) reported no main expectancy effect in a middle-class elementary school in the Midwest. Girl bloomers gained less than controls while boy bloomers gained more. Conn, Edwards, Rosenthal, and Crowne (1968) studied 258 children in a grammar school (grades 1–6). IQ was the main dependent variable. "There were no clear expectancy effects" (Baker and Crist 1971:50), but there were differences in sensitivity to emotional communication, especially among boys. Anderson and Rosenthal (1968) studied twenty-eight retarded boys. There was no significant IQ gain as a result of expectancy. There was no evidence of main effects in Rosenthal, Baratz, and Hall (1974). Pygmalion is a textbook case of cherry-picking the most favorable results for publication.

In addition to Rosenthal's own work, there are several meta-analyses that summarize the work of other researchers on teacher expectations. Rosenthal cites these to his advantage. The hundreds of studies that Rosenthal refers to as supportive of the "Pygmalion effect" are not replications of the IQ study, but studies of how expectation effects color the atmosphere, feedback, input practices, and output opportunities in settings where people have been led to believe that they will interact with others who are more compatible, smarter, more interesting, and so on. These studies record *experimenter effects*, not IQ shifts. Two meta-analyses, however, do review the latter sort of studies. A short report by Smith in 1980 suggested that the correlation between teacher expectancy and pupil IQ was $r = 0.08$. Although Smith's study is cited favorably by Rosenthal (1987), in point of fact Smith (1980:54) concluded that pupils' intellectual ability was "minimally affected" by manipulated expectations. Raudenbush's 1984 meta-analysis had a more sensitive focus: examining the magnitude of the IQ affect while controlling for *prior* acquaintance of teachers with pupils. Raudenbush discovered that the magnitude of the effect was greatest where prior contact was smallest. Based on the meta-analyses, Rosenthal suggested that "the educational self-fulfilling prophecy (Merton 1948) has now been well established." He based this conclusion on eighteen studies and claimed "the effect of teacher expectations were significant for his full set of 18 studies" (1987:39). But Raudenbush deduced that the correlation was small (r = 0.15) on the basis of the seven studies that most credibly met the criterion of little prior acquaintance (i.e., less than a week). With respect to the eighteen studies, in fact, four tests of association were explored. The only one that proved non-significant was the one in which a control was employed for sample sizes. "Larger sample sizes tend to produce smaller effects" (Raudenbush 1984, cited in Wineburg 1987b:43).

A final point from Wineburg: Raudenbush's mean effect size for the expectancy–IQ link is 0.11 (standard deviation = 0.20),

> but as any introductory student knows, the mean is notoriously sensitive to extreme values when a distribution is skewed. The median effect size of Raudenbush's 18 studies is but .035; ten studies yielded positive difference and eight yielded negative differences. What kind of phenomenon is it when nearly half the attempts to produce it yield results in the wrong direction?"
>
> (1987b:43)

So, even examining the most relevant cases, the Pygmalion effect is precariously close to zero. For Wineburg, the Pygmalion study was "the self-fulfillment of the self-fulfilling prophecy." And it led him to ask: "does research count in the lives of behavioral scientists, teachers and children? If not, we might as well close up shop and refer all correspondence to Family Circle" (1987b:43). In addition to these substantive problems, there were the ethical issues. It has already been noted that there was no informed consent from parents or teachers. After the fact, the teachers appeared to have been debriefed (Ellison 2015), but no one explained to the students or parents why students were maintained within grade levels for a period of two years despite dramatic changes in their academic ability.

The legal legacy

Even lousy social science can have powerful political and legal consequences. Pygmalion has been cited in support of actions to force busing in two American jurisdictions. Busing was justified on the need to counter racist attitudes that damage minority students. Pygmalion was used to further social objectives and progressive public policies in spite of its academic shortcomings. From the judgment of Judge Wright in *Hobsen v. Hansen* 269 F. Supp. 401 (1967), a case that supported forced busing to integrate multiracial mixing in the U.S. schools we read:

> Studies have found that a teacher will commonly tend to underestimate the abilities of disadvantaged children and will treat them accordingly – in the daily classroom routine, in grading, and in evaluating these students' likelihood of achieving in the future. The horrible consequence of a teacher's low expectation is that it tends to be a self-fulfilling prophecy. The unfortunate students, treated as if they are subnormal, come to accept as a fact that they ARE subnormal.
>
> (*Hobson v. Hansen,* p. 484)

Wineburg commented in the *Educational Researcher*:

> To substantiate these claims, Judge Wright cited two studies: one by Clark (1963), which presented no data directly bearing on the self-fulfilling prophecy, and an edited chapter on the Pygmalion study. But unbeknownst to him, Pygmalion dealt with the overestimation, not underestimation, of children's abilities. Moreover, it presented no observational data of teachers and students, so there was no information on how teacher's "treated" students. Further, no interviews were conducted with students to see whether they accepted their "subnormal" status. Although all the points raised by Judge Wright may in fact be true, Pygmalion did not provide the evidence.
>
> (1987b:33)

Conclusion: social psychology and social engineering

It is not an exaggeration to say that millions of students in North America and Europe have been exposed to the conventional views of Pygmalion and the Hawthorne effect. Many consumers of this information would have been heartened by the potential for improving society through following the "lessons" supposedly learned at General Electric's Hawthorne manufacturing plant in Cicero, Illinois, and in the Spruce School in Los Angeles. They have been seduced by what turn out to be scientific myths. Few would reflect on the fact that the case for the Hawthorne effect was based on a mere five workers (seven if we count the replacements), and, in the teacher expectation study, on just nineteen students. Their methodological flaws should have condemned these studies to the dustbin of scientific history. But the potential for improving society by the social engineering ideas of psychology overshadowed the evidence on which they were based. These apparent miracles of science enchanted generations of students, professors, as well as the general public. Unfortunately, the scientific progress attributed to these studies was an illusion. The lesson suggested by these cases is that their moral appeal has more than compensated for their total empirical bankruptcy.

Contemporary scientific writing on the Hawthorne Effect remains lively but it is no more conclusive today than it ever has been. In their meta-analysis of recent research McCambridge, Witton, and Elbourne (2014) reviewed nineteen recent studies of the Hawthorne Effect (eight randomized experiments, five quasi-experimental field studies and six observational studies). The objective of the review was to determine whether there was any evidence of an effect and to estimate its size based on evidence available across several scientific disciplines. However, they note that the concept has been used in the recent literature "without any necessary connection to the original studies" (2014:268). This reflects the "illusion of familiarity" discussed earlier. Whatever it originally referred to, the effect is now taken to refer to the behavioral consequences of being observed. It assumes that awareness of being observed primes the person to the observer's expectations, leading to changes that conform to those expectations. This is not limited to industrial productivity. The priming or communication of the Hawthorne-type expectations could be done, for example, through interviewing voters before an election, administering a questionnaire before screening of patients for cancer treatment, being aware that one is in a group of treatment participants for tuberculosis therapy, or announcement of a study in a memo to paramedics about their record-keeping. The measured outcomes were whether people voted, the uptaking of screening opportunities for cancer detection, the retention of patients in tuberculosis treatments, and the recording of medications, allergies, and medical histories by paramedics. Of the nineteen studies, twelve provided at least some evidence of an effect, however small. What do the effects mean? According to McCambridge et al.,

> [t[]hese data suggest that the size of any effects of health-care practitioners being observed or being aware of being studied probably very much depends on what exactly they are doing … the effect, if it exists, is highly contingent on task and context.
>
> (2014:275)

They go on to say that "there is no single Hawthorne Effect" (2014:276), To be more specific, persons interviewed before an election are a bit more inclined to vote. Persons who join a group for tuberculosis treatment are inclined to drop out at a little bit lower rate than those who are not attached to a group. Paramedics who are notified in a memo that they are being studied provide a somewhat more detailed record of their interventions than otherwise. In short, persons who believe that their behaviors are being monitored sometimes change those behaviors. This amounts to concluding that individual behaviors are socially sensitive. Sometimes. Conceptually, this is stunningly vague. The authors note "[a]s the Hawthorne effect construct has not successfully led to important research advances in this area over a period of 60 years, new concepts are needed" (2014:268). What the meta-analysis reveals is the enormous heterogeneity of activities that have been described as illustrations of the Hawthorne concept outside of the context of worker productivity. New concepts are needed because the metastasis of the concept beyond the original focus on worker productivity has trivialized our knowledge by concluding that human behavior is sometimes shaped by the expectations of others. That is common sense.

However, within industrial psychology, the concept survives intact. Writing in *Industrial Management*, Chris Porter (2012) notes: "More than eight decades after the initial experiments, the Hawthorne effect remains with us today … Effective supervision can push employees to greater heights without the need for expensive technological solutions." The *concept* may remain with us today, but *evidence* for its revolutionary impact on productivity does not. In management schools, there is a flourishing literature of the imperative of communicating high expectations in respect of employee morale, not because this practice has reliable outcomes, but because the Hawthorne myth has made this a professional standard.

Similarly, in education, the Pygmalion myth is honored because it has implications for the professional conduct of educators *vis-à-vis* their students. Timmermans, Rubie-Davies, and Rjosk (2018) reviewed the state of the art in teacher expectations in the five decades since the publication of Pygmalion in *Educational Research and Evaluation*. The research indicates that teachers have a relatively accurate understanding of their students' abilities, but that they sometimes arbitrarily favor some students over others, that they seem to favor students from more affluent families, and that they have lower expectations for students with special needs. Evidence for teacher beliefs in gender differences in their students' abilities in mathematics versus language is not consistently supported in the literature. Also, there is a significant relationship between teacher expectations and the performance of their students, but this does not mean that an artificial change in the teacher expectations produces a bump in student performance. In fact, in their assessment of the relationship between "teachers' expectations and student achievement", De Boer, Timmermans, and van der Werf (2018:183), expressly excluded Pygmalion-like interventions "as they did not aim to evoke a sustainable change in teacher expectations, nor did they have direct applicability to regular classrooms." Such strategies as prioritizing feedback between teachers and students have a modest contribution to student performance (d= 0.43, Hattie 2009:13), but random creation of expectations as in Pygmalion have a reportedly tiny effect ($r = 0.1$, Timmermans, Rubie-Davies, and Rjosk 2018:91).

The contemporary educational establishment is not invested in the claims of magical outcomes in student achievement from changes in teacher expectations. Emphasis is more on a conscientious communication of a positive relationship with students to avoid compromising their development. In both business management and in education, the idea that there is an easy path to massive improvements in worker productivity and student IQ is simply wishful thinking.

Notes

[1] Marion Gross Sobol says: "The effect has been referred to as the 'Hawthorne effect' by Lazarsfeld in his article 'Repeated Interviews as a tool for studying changes in opinion and their causes'," (1959: footnote 1) in the *American Statistical Association Bulletin* 2:3–7 (1941).

2 Parsons does not provide a test of significance in output, preferring to report instead graphs that capture evidence of increases in “total output” for 1927–1929, followed by a decline in hours worked and total output from 1930 to 1932 (Parsons 1974:926). In comparing Parsons and Jones, we find Parsons noting the increase in output while Jones’ test of the effects of remuneration is non-significant because there is virtually no variability in the method of payment over his time series. However, everyone agrees that there was some increase, especially in the first thirteen periods of the study. Carey (1967:405–8) puts the increase at about 15%.

3 By this time the Pygmalion study had already been under way for two years.

4 Elashoff and Snow (1971:158) discovered that there were already significant IQ differences between the control and experimental subjects from the start (4.9 IQ points higher for verbal and 13.2 points higher for reasoning IQ). Only positive expectations were created.

5 See Rosenthal and Jacobsen (1968) Figure and Table 9.1. The references in subsequent notes refer to figures and tables in the original Rosenthal and Jacobson 1968 book.

6 See Figure and Table 7.1.

7 See Figure 7.2.

8 See Figure/Table 9.2.

9 See Figure/Table 8.1.

10 See Table 8.5.

11 See Tables 7.3 and 7.4.

7 A guide to the myth of media effects

Introduction

It is difficult to determine when fears about the adverse affects of mass media became a preoccupation of the respectable classes, and motivated attempts to bring the worrisome elements of popular fiction under the control of the state and the courts. In his history of pornography, Walter Kendrick (1988), records how the 18th-century excavation of the ruins of Pompeii, buried in 79 A.D. by the eruption of Mount Vesuvius, brought to light household statues and paintings from ancient Roman culture that revealed an extraordinary sexual frankness. The Naples museum accessioned frescoes of nude females, couples making love, satyrs having sex with goats, phalluses in relief in paving stones and on the walls of houses, statues with oversized penises and phallic ornaments in household appliances and birdfeeders, to name a few items. In addition, Pompeii apparently had hundreds of brothels but the erotic artifacts were found throughout the city.[1] Such vivid sexual representation was found so offensive in Enlightenment Europe that the materials were housed in a secret museum in Naples that restricted public access. The very catalog was considered X-rated and access was confined to the male ranks of the privileged classes. What Kendrick calls the "pre-pornographic" culture began to unwind in the 19th century. Puritanical sexual inhibitions came under pressure with the spread of literacy in the 19th century. Popular fiction exposed readers to the negative influence of permissive writing. "Genteel society" objected to the publication of Mark Twain's *Adventures of Huckleberry Finn* in 1885, and the book was banned from the Concord Library in Massachusetts as "trash suitable only for the ghetto." There are many tokens of this change. In the late 1930s, the appearance of comic books was greeted by alarm. Many U.S. states as well as Canada, the UK, Australia, and several European countries attempted to ban the comics under obscenity laws. According to Frederic Wertham, a New York psychiatrist, the provocative "comic" covers, the adult crime themes, and the celebration of violence and crime were believed to promote juvenile delinquency, racism, and homosexuality. His book, *Seduction of the Innocents* (1954), and his coverage in popular magazines struck a chord with the public. Here was a psychiatric authority who shared the public's misgivings about the influence of perverse literature. He wrote

> we do not maintain that comic books automatically cause delinquency in every child reader. But we found that comic-book reading was a distinct influencing factor in the case of every single delinquent or disturbed child we studied.
>
> (quoted in Crist 1948:22)

A public inquiry into the causes of delinquency in the U.S. in the early 1950s chaired by Senator Estes Kefauver raised questions about the influence of mass media on youthful misconduct. Questions of imitation of criminal behavior had been raised earlier in some of the Payne Foundation studies of motion pictures in the 1930s and 1940s. Paul Lazarsfeld had studied the influence of radio on political opinions during the same period, although he found little evidence for significant shifts in political opinion related to campaign speeches (see Hovland 1959). However, U.S. academic funding of mass media effects, especially television, was in its infancy until the early 1960s. The work of Bandura, Ross, and Ross (1963) raised questions about the vulnerability of youth to messages of violence in children's programming, including the popular Saturday morning cartoons. Bandura (1973) advanced a model of influence known as "social learning theory," in which people could acquire new behaviors through vicarious experience, that is, seeing other people benefiting from a specific behavior and mimicking it with expectations of gaining a similar benefit. The 1960s saw the emergence of laboratory studies of imitative violence, as well as long-term field studies of the correlation between violent media exposure and aggressive behavior. Interest in the question of media effects in America grew as crime rates exploded throughout the 1950s, 1960s, and 1970s.

In 1972, the U.S. Surgeon General issued a report on the effects of television violence that cautiously expressed some concerns about the inadvertent effect of violent entertainment on viewers. Its conclusions were couched in highly conditional language: for example, television "can" induce short-term mimicry in children, it "can" instigate an increase in aggressive acts; however, the evidence "does not warrant the conclusion that television violence has a uniformly adverse effect nor the conclusion that it has an adverse effect on the majority of children" (National Institute of Mental Health 1982, cited in Liebert and Sprafkin 1988:113). This cautious publication appeared in the golden age of media effects research. At the instigation of mass media researchers, a new report was prepared a decade later under the auspices of the National Institute of Mental Health: *Television and Behavior: Ten Years of Scientific Progress and Implications for the Eighties.* The 1982 report updated the empirical findings, and concluded that the evidence supported inferences of a *causal* link between exposure to violent television and aggressive behavior among viewers, suggesting that there was a wide consensus among social scientists about this fact, and that the conclusions were based on the "convergence" of different kinds of evidence, none decisive on its own, but convincing when taken together.

> In the 20 years since the publication of the Surgeon General's report, research into the TV violence issue has burgeoned. Laboratory experiments

> continue to provide evidence of a causal relationship between violence viewing and aggression. The results of nonexperimental field studies support the same conclusion … The majority of new investigations suggest that viewing violent entertainment can increase aggression and cultivate the perception that the world is a mean and scary place.
>
> Liebert and Sprafkin (1988)

The NIMH report created a lot of public and academic debate. Studies of the effects of television violence on children gave way to investigations of the impact of pornography on male readers and viewers, using many of the same experimental protocols employed in the television research. Theories of media effects appeared in the court system, sometimes as evidence in support of defenses of temporary media-induced insanity, sometimes in cases where victims of violence sought compensation from broadcasters for attacks resulting from media imitation. In the case of pornography, evidence from experimental social psychologists was influential in challenging the common law of obscenity.

After all this time and effort, one would think that psychology had come to some firm conclusions about the way violent media influence human behavior, and that regulatory action could be developed on evidence-based policies. But that is not what occurred. In what follows, I examine the preoccupation of media studies with violence. I then examine whether the social learning theory is a truly distinct form of learning. We shall explore some peculiarities of the logic of experimental social psychology, the difference between statistical significance and effect sizes, and the policy significance of decontextualized effect studies. I will explore the extra-scientific incentives that underlie the media effects academy. And end with a review of the most recent controversy over violent video games and attempts to regulate them legally. We begin with the preoccupation with media violence.

Social learning theory, TV, and the spectre of violence

Psychologists are proud to point out that the field is based on three great models of learning. First, there is the classical conditioning model based on experiments with Pavlov's dogs, where he produced salivation by pairing an unconditioned stimulus (a bell) with a conditioned stimulus (food), showing that the animals learned to salivate at the sound of the bell. Second, there is operant conditioning, or trial and error learning, based on (among other things) Skinner's studies of rats and T-mazes, and the differential rewards attached to choosing correct pathways, rewards that accelerated animal responses. In the late 1950s, experimental social psychologists sought out a third explanation of learned behavior peculiar to humans, and reflective of their sophisticated cognitive abilities. Bandura's *social* learning theory was the chief theoretical basis for laboratory studies of the behavioral and attitudinal changes attributed to the new electronic media. However, it had a number of peculiarities that ultimately would impede its success. In retrospect, the first was the remarkable narrowness in what it examined. It was preoccupied with aggression and selected stimulus programs from *Batman* to the *Road Runner* cartoons, as though aggression was the main "lesson" of such programs and as though this was the only effect worth noting. Television violence was the theory, aggression was the practice. Aggression became a proxy for all that was problematic with youthful behavior. Nobody seemed to wonder whether films that portrayed theft, prostitution, or narcotics encouraged viewers to steal, prostitute, or get stoned. Did comedies make viewers inclined to tell funny stories and act like comedians? Did *Raiders of the Lost Ark* and *The Temple of Doom* encourage viewers to study archaeology? Did *Caddy Shack* encourage viewers to take up golf?

The virtually exclusive focus on aggression has been remarkably one-dimensional. This has gone hand in hand with a presupposition about the peculiar vulnerability of children. The implication is that social learning occurs primarily in childhood. The famous longitudinal studies of the lagged effect of early childhood TV viewing habits on aggression (Huesmann, Eron, Lefkowitz, and Walder 1973, 1984) ten and twenty years later assume an age-graded developmental model in which violent media exposure creates persistent antisocial behaviors, and that these are determined by media exposure in "the early window" of life.

Of course, this model is contradicted when the same media effects paradigm is applied to pornography. There are reports that suggest that even extremely short exposures to sexually violent, highly arousing pictures can foster rape fantasies and calloused attitudes in otherwise normal men (Malamuth and Check 1981). Men, like children, appear to be extremely vulnerable to the media, especially sexually violent media exposure. But there is no supposition here of an age-contingent impact, since the studies of pornography effects deal exclusively with adults, that is, persons whose dispositions one would have thought were already stable, having been laid down in childhood. When one sees the research agenda of the media effects experts in the 1960s, 1970s, and 1980s, it is hard to escape the proposition that the effects of greatest interest were decided *a priori* in the world of respectable fears about the vulnerability of children and women, and the apparent invasiveness of the new technologies. For every one study on the positive effects of children's programming on literacy, there were twenty-five studies on the consequences of sex and violence. However, even the evidence for the positive effects of *Sesame Street* were probably mediated by child–parent interaction in the context of viewing (Wurtzel and Lometti 1984:36).

In 1971, Leonard Berkowitz wrote an influential comment in *Psychology Today* in which he contrasted the 1970 Presidential Commission on Obscenity and Pornography and the Presidential Commission on the Causes and Prevention of Violence. "Sex and Violence: We can't have it both ways." The former concluded that exposure to sexually explicit fiction was unassociated with harmful behavioral consequences, while the latter found that violence on TV promoted violence in everyday life and reinforced attitudes conducive to violence. The former report called for a liberal policy towards sexual fiction, the latter called for elimination of violence in children's cartoons and a reduction in violent programming more generally. While on the surface these looked like radically inconsistent understandings of media effects, we should not overlook the fact that the violence commission was looking at the vulnerability of children and youth, while the pornography commission was looking at adults.

Social learning and common sense

A second point is that the social learning model appears to be quite a modest intellectual perspective, when we move away from specific technologies (television) and specific prohibited behaviors (aggression) and when we conceive of it in more general terms. Note that there is nothing in the theory that would establish that persons are *peculiarly* vulnerable to digital information, as opposed to other sources of information – parents, siblings, peers, school teachers, newspapers, books, neighbors, moral teaching, etc. But social learning theory cut its teeth on television and the subsequent video technology without reference to how social learning worked in all human history prior to the intrusion of the national networks into family life, and how it continued to work in everyday life outside television. In other words, the very idea of social learning only emerged with the rise of television. Furthermore, the intellectual study of social learning co-appeared with a specific social agenda: its long-standing interests in censoring children's programming, a position that amounts to promoting certain social values in the name of "public health." The social agenda of regulating television was premised on the understanding that it had become one of the most important sources of learning, that it was the source of *decisive* social models and, hence, one of the most important determinants of behavior, including deviant behavior. This was a *premise* from the very start, and one that the experiments were designed to demonstrate. The early Bobo doll studies were designed to showcase how children learned bad habits from watching cartoons. The fact that the Bobo doll was *designed* to be punched, and that some forms of punching and roughhousing in the context of male play can be wholesome behavior were overlooked. Like other studies in the golden age of experimentation, the lesson was allegorical. Subtract the element of play from the subjects' response, equate the Bobo doll to another innocent child, and, *voilà*, our worst fears were realized.

There is a definite moral cast in the attitudes of psychologists to popular television dramas and cartoons in the 1960s. They assume that the violence in *Batman*, *Superman*, and cowboy fiction is rewarded, and that the viewers are converted to the dark side as a consequence. But the vast majority of so-called violent fiction results in the punishment of the wicked and the reward of the virtuous (Fowles 1999). This is the cathartic attraction of all popular fiction, a point consistently missed by those who believe that catharsis in fiction is only interesting if it alters behavior afterwards and somehow lets the viewer "blow off steam." The catharsis I refer to is the identification with the hero, the suspense experienced as the hero faces danger, our fear and contempt of the villains, our anxieties as the plot goes their way, and our relief as the cavalry rides in to restore what used to be referred to as "truth, justice, and the American way." Without catharsis, fiction would be totally lacking any dramatic attraction. Why would we presuppose that human viewers would miss the almost clichéd moral nature of popular entertainment, the Punch and Judy portrayal of good and evil? In my view, the failure to grasp this point has been a fatal flaw in social learning theory, and has led researchers to overlook the situational impact of even short-term media exposure on emotional volatility in the laboratory, something that is frequently misinterpreted as imitation. We shall return to this point later, since it addresses a major theoretical inconsistency in understanding the meaning of the apparently high impact of media on behavior in laboratory studies of aggression.

Is social learning theory distinctive?

What is the actual mechanism (or mechanisms) that contributes to social learning? Psychologists write at times as though changes in behavior occur below the threshold of consciousness. When social learning occurs, are we simply talking about absorbing cultural images as though they were normal, that is, unnoticed? In other words, we grow up in the Shire and think as Hobbits or we grow up in Rivendel and think as Elves (Tolkien 1966). The worldviews are radically different but experienced as natural. Stereotypes are probably communicated this way. Indeed, there was some concern among psychologists that television shows like *Amos and Andy* portrayed African-Americans in racial stereotypes. The same point is made in the analysis of pornography as an insult to the status of women. But the concept of *selective exposure* to certain worldviews is not a specific mechanism of learning, unless we acquire conditioned responses to categories of people in the same process. In that case, this conceptualization of social learning amounts to classical conditioning. In other words, my reaction to a specific class of people (for example, a member of a racial, religious, or linguistic group) is conditioned by stereotypes or idealizations (favorable or unfavorable) reinforced for such groups in the media.

A second issue is whether the observed behavior that is mimicked must result in a reward of which we vicariously approve and pursue for our own benefit. We mimic a colleague's behavior because we see that it resulted in certain benefits. Does this happen subconsciously? Or would we be aware of the consequences of action and consciously make similar choices? Is that not a case of "judgment," that is, choosing wisely through generalization from a stimulus in our own everyday experiences? In that case, this conceptualization of social learning amounts to operant conditioning. I follow my mentor's advice because I expect to benefit from it – generalizing from his or her case to my own. Following this line of reasoning, social learning is not a distinctive form of learning. Indeed, it is arguably trivial or obvious. The argument from social learning theory amounts to the observation that we take our own culture for granted (selective exposure leading to classical conditioning) and/or we maximize our utilities as classical economics suggests (trial and error). There is nothing new or mysterious here. Perhaps the paradigm's focus on children makes such selective exposure (classic conditioning) on the one hand, and self-reflection (operant conditioning) on the other less than obvious, since we assume children are naïve and unreflective. Again, that would be an indication that social learning is premised on immaturity, or a specific theory of development, as noted earlier.

Is there another approach to social learning that tacitly acknowledges the power of fiction to create arousal? When we observe that children may be acting more immaturely or impulsively in the aftermath of provocative images, are we not registering the cathartic, stimulating (i.e., plain "entertaining") effects of drama? When we "act out" in the aftermath of exposure to a "violent" cartoon, have we somehow been seduced by a subconscious mechanism of which we are only dimly aware? Is there some magical effect that sweeps us up into activities that would be terminated on a moment of mature self-reflection and/or social control? And do our individual careers evolve by such acts of magic? In either case, social learning theory would be really very novel, non-trivial, and provocative. I think that something like that does occur and that it is tied to catharsis (i.e., arousal). We shall get to the magic in due course. We began this discussion by questioning whether there really is strong consensus about harmful effects of violent fiction.

Where is the consensus about media effects?

What does the record show regarding consensus and why is that important? In 1995, Leonard Eron told a U.S. Senate hearing that "the scientific debate is over" about the harmful effects of violent television programming, that the relationship was a settled scientific fact (Fowles 1999:20). The record suggests otherwise, both in the area of violent television and pornography. Psychologists who thought television was creating negative impacts on younger viewers besmirched the motives of the television industry experts who contested the evidence of harmful effects, as though academic researchers were impervious to their own interests. The gold ribbon panel that assembled the NIMH document wrote that "it would be no exaggeration to compare [the] attempt by the television industry [to contest evidence of harm] to the stubborn public position of the tobacco industry on the scientific evidence about smoking and health" (Chaffee et al. 1984:30–1). The parallels strike me as hugely self-serving, since they imply that people who work for universities have fewer career interests and ideological preferences than those in industry, a position difficult to sustain if we reflect on the previous chapters. In the case of tobacco, animal studies could be used with nicotine products to induce cancers in mice, providing evidence of a strong, direct effect. Laboratory studies of human violence tended to employ tests based on analogies to violence that were typically short-lived and metaphorical. Evidence from field studies was inconsistent and contradictory, and the zero order correlations (r) between media exposure and subsequent behavior were typically weak (i.e., $r = 0.15$). Failures to replicate the presumably well-established findings from the laboratory have been common (Linz and

Donnerstein 1988; Boyd and Brannigan 1991; Fisher and Grenier 1994). Also, the social nature of aggression studied in the laboratory and violence in everyday life are typically worlds apart. A methodological note is in order here regarding the experimental approach. Laboratory studies of aggression were conducted primarily within a research protocol called "the Buss paradigm."

The Buss paradigm was developed to explore the causes of aggression. Subjects are put into a state of high emotional excitement, typically by being insulted by a confederate (a person working for the experimenter but pretending to be just another subject). Subjects are then assigned to some kind of treatment condition (for example, one of several different kinds of video), and finally, in a supposedly unrelated third phase of the experiment, are asked to administer shocks to the person who earlier had insulted them. The experiment picks up the media effect of the intervening treatment usually only in the presence of the state of high arousal. Berkowitz puts it this way: "The observer will exhibit the highest aggressive reactions if he is emotionally aroused at the time, believes his aggressive actions will have favorable rather than unfavorable consequences, and thinks the observed victims had deserved the injury inflicted on them" (1971:18). So, the evidence of harm is somewhat oblique. As in the Milgram paradigm, the subjects are encouraged to administer shocks to teach them a task, as though this were legitimate ("favorable"), and then their behaviors are equated with giving "injury."

The limited utility of the laboratory evidence for the impact of media on aggression is suggested by the treatment of this subject in the classic textbooks on criminology. Wilson and Herrnstein (1985:337ff.) review the case for television and the mass media impact on crime in *Crime and Human Nature.* Of the field studies, they say, "The best studies come to contradictory conclusions, and even when all doubts are resolved in favor of a causal effect, they account for only 'trivial proportions' of individual differences in aggression" (1985:346). When experts assembled by the National Research Council reviewed the 1982 NIMH report, they concluded that televised violence "may" be related to aggression, "but the magnitude of the relationship is small and the meaning of aggression is unclear" (1985:353). "Even giving to existing research the most generous interpretation, viewing televised violence cannot explain more than a very small proportion of the variation in aggressive acts among young persons" (1985:353). Similarly, Kaplan and Singer argued, following a review of the literature, that "this research has failed to demonstrate that TV appreciably affects aggression in our daily lives" (1976:62). Feshbach and Singer reported no evidence from field studies that violent fantasy programming increased aggression. In fact, there was some evidence "that exposure to aggressive content on television seems to reduce or control the expression of aggression in aggressive boys from low socioeconomic backgrounds" (1971:145). In Freedman's review of the field studies, he reported that "not one study produced strong consistent results, and most produced a substantial number of negative findings" (1988:158; see also Freedman 1984, 1986). David Gauntlett reported "the search for 'direct' effects of television on behavior is over: Every effort has been made and they simply cannot be found" (1995:120). Gadow and Sprafkin: "The findings from the field experiments offer little support for the media aggression hypothesis" (1989:404). Attempts to replicate abroad earlier work conducted in America by Huesmann et al. drew the same disappointing reactions. Wiegman, Kuttschreuter, and Baarda reported from Australia: "These results give no support for the hypothesis that television violence viewing will, in the long term, contribute to a higher level of aggression in children" (1992:155).

This shows up one of the peculiar methodological directions resulting from a nearly exclusive reliance on experimentation and quasi-experimentation. The subject matter of research is more focused on the predictor (the causes) than the outcome (the effects). This permits the experimenter to hold onto a cause, no matter how substantively trivial, and to remain ignorant of the main contours of the phenomenon of interest that it effects. The media effects research shed light on the programs but told us virtually nothing about the phenomenon of violence in everyday life. Rather than asking what are the major causes of violence, and what are the recurrent contributions of gender, age, individual differences, and class, all our time is devoted to a single predictor – the media. As many commentators have pointed out, this probably reflects the value orientation of the researchers.

> Social scientists tend to abhor violence and dislike much of popular culture; it is only natural, therefore, that when the public worries about what television may do to their children, especially when there is a rising level of violence in society, scholars would concentrate their efforts on showing how televised violence … increases the violence of television viewers, especially children.
>
> (Wilson and Herrnstein 1985:339)

One of the costs of this approach is that it is unclear how much effect censorship policies could be expected to have even if they were adopted. If the correlation between media and subsequent behavior is 0.1 to 0.2, the explained variance, adjusted for all the other major predictors of aggression, would be about 1–4%. Liebert and Sprafkin argue in defense of the monocausal analysis as follows: "Researchers have said that TV violence is *a* cause of aggressiveness, not that it is *the* cause of aggressiveness. There is no *one,* single cause of any social behavior" (1988:161, emphasis in original). This is true, but without some knowledge of the other leading causes, it is impossible to identify its *relative* importance. And without that, we cannot determine whether something that is statistically significant has any social significance. It is noteworthy that in the Surgeon General's report on *Youth Violence* (Satcher 1999), reference to media effects is virtually absent, in spite of the views of some medical experts that violent programming is "the number one public health issue" responsible, in the estimates of Brandon Centerwall (1993:63), for 10,000 homicides each year in the United States.

The issue of media effects is discussed in a second leading criminology textbook, *A General Theory of Crime.* In their review of psychological positivism, and its preoccupation with aggression, Gottfredson and Hirschi (1990:67ff.) note that psychologists sometimes stumble into the larger world of misconduct outside aggression. Huesmann, Eron, Lefkowitz, and Walder define aggression as:

> an act that injures or irritates another person … but makes no distinction between accidental and instrumental aggression or between socially acceptable and antisocial aggression. The assumption is that there is a response class, aggression, that can include a variety of behaviors, exhibited in numerous situations, all of which result in injury or irritation to another person. This category includes both hitting and hurting behaviors, whether or not these behaviors are reinforced by pain cues from the victim or target person. This category also includes injury to or theft of property.
>
> (cited in Eron 1987:435)

This definition conflates "socially acceptable and antisocial aggression" – but of what policy utility is that and how can it "result in injury"? The definition also arbitrarily includes property crimes. However, the inclusion of such crimes would make sense if the "response class" is a general tendency toward dysfunctional or impulsive behavior. That comes close to the criminological understanding, as we shall see. In the "Rip Van Winkle study," as it has become known, Huesmann, Eron, Lefkowitz, and Walder examined 875 subjects in grade three classes in upper New York State, starting in 1960, identifying television viewing habits and peer-nominated levels of aggression. Mothers were asked to identify the children's three most popular television programs, which were then classified as either violent or non-violent. Ten years later, and again twenty-two years later, the subjects were followed up to determine their viewing habits as well as their criminal involvement. The findings suggested that early signs of aggression, especially in males, predicted patterns of aggression later in life, that the most aggressive children watched the most television, and that aggression was also related to lower IQ. Eron attributed the aggression to "continued television violence viewing" (1987:440) (although continued viewing was not observed). The study did not establish baseline IQ and aggression scores which may have influenced which types of programming the children chose to watch.

From their perspective, Gottfredson and Hirschi point out that television viewing at age eight would equally well predict "theft, motor-vehicle accidents, trivial nonviolent offending, drug consumption, and employment instability, behaviors hard to attribute to the number of shootings or fistfights watched on television twenty years previously" (1990:69). Indeed, Eron reports that childhood aggression in fact predicted "social failure, psychopathology, aggression and low educational and occupational success" (1987:440) twenty-two years later. Why? All kinds of dysfunctional behaviors tend to cluster in the same persons. Individuals who are "aggressive" do not tend to specialize in aggressiveness but exhibit a short temporal horizon or impulsiveness across whole "response classes." Gottfredson and Hirschi: "It therefore seems unlikely that the specific content of television programming viewed at age eight could contribute independently to subsequent levels of aggression" (1990:69) This is because they are all an expression of the same stable traits. This could explain why media viewing habits could be correlated with a range of dysfunctional behaviors without being an important determinant of them. Persons with a high tolerance of risk are not disturbed by provocative excitement, so the co-appearances of a high-exposure threshold and a high-risk behavioral threshold are predictable correlations. The traits that Huesmann et al. documented showed remarkable stability over the life course. "What is not arguable is that aggressive behavior, however engendered, once established, remains remarkably stable across time, situations and even generations within a family" (1984:1133). This was consistent with the published report of Olweus (1979:866), who showed impressive continuity in individual traits over the life cycle; over a ten-year period, the estimated coefficients were on the order of 0.60 for aggression and 0.70 for intelligence.

Causal testing versus population processes

There is one further point that must be mentioned regarding the utility of media effects in the context of criminology versus experimental psychology. As Berkowitz and Donnerstein (1982) argue, the theoretical purpose of an experiment may not be to determine a population estimate of some effect. Indeed, for some purposes it might be totally appropriate in order to test a specific causal relationship that the laboratory setting has little "mundane realism," that is, little resemblance to familiar situations in everyday life. For example, the Buss paradigm may be attractive for teasing out the relative impact of an *explicit* film versus a *sexually aggressive* film. The fact that the subjects begin in a state of high emotional discharge, and that the experiment forces retaliatory aggression without permitting the subjects an opportunity to calm down, may permit the estimation of the marginal differences in behavior resulting from different stimuli when all the other factors are held constant and/or when all the common restraints on aggression are eliminated. As a result, this situation may not correspond to anything found in everyday life. It may correspond to natural situations by degree, but, in some cases, it may be something found only in the laboratory. Henshel (1980) makes a case for this when he points out, for example, that temperatures of zero degrees Kelvin (and the effect of absolute cold on the electrical properties of magnets) are not found in nature, but can be created in a laboratory. Similarly, the teaching of American sign language to primates may permit scientists to learn something about primate cognitive abilities but it is not natural behavior. Berkowitz and Donnerstein recognize this possibility when they say "we are not insisting that the laboratory findings are necessarily generalizable to the world outside" (1982:255). There is a recognition that the experimentalist oftentimes exchanges ecological validity for causal control. What is difficult to determine is the extent to which the controlled world of the laboratory actually helps us understand the magnitude of manipulations when we want to extrapolate back from the controlled environment to everyday life.

> Even if we confine ourselves to psychological influences, the laboratory setting is not necessarily representative of the social world within which many people act. As a consequence we cannot use laboratory findings to estimate the likelihood that a certain class of responses will occur in naturalistic situations. Suppose that 60% of the subjects in a sample exhibit heightened aggressiveness over some baseline level when a weapon is present. Even if these people were representative of the persons in a larger population, we could not say, of course, that 60% of this broader group would react in the same way in a more realistic situation. Experiments are not conducted to yield such an estimate.
>
> (Berkowitz and Donnerstein 1982:255)

Therefore, discovering a causal relationship (which may be positive or negative) and a population estimate of its magnitude are quite separate objectives. However, there is another matter raised in the title – "external validity is more than skin deep." Berkowitz and Donnerstein stress that, despite the artificiality of the learning situation, the experiment is valid because the subjects in the Buss paradigm experience an intense desire for revenge, and they administer shocks believing that they are hurting someone who deserves to be punished for insulting them. In other words, the experiment has intense realism for participants that is more than skin deep. They "mean" it. However, if this is the real point of distinguishing causal testing and a population modeling, two problems arise. First, because experiments are not based on a sampling frame, it is never possible to provide population estimates of anything with confidence. Second, the claim that the artificial manipulations still retain "realism" amounts to a claim that the *internal* validity of the experiment is sound, and that the definition of the situation (insult and subjective provocation) is credible to the subjects even if the tasks that produce such reactions are unfamiliar: that is, in this case, that the participants' conduct was sincere, and not just role-playing or following demand characteristics. But that is an empirical question the answer to which would turn on a validity check. "Appropriate questioning is vital to insure that the participants have interpreted the experimental treatments in the desired way" (Berkowitz and Donnerstein 1982:255). Even if the subjects defined the situation as serious, the experimenter is claiming external or ecological validity without knowing the probable ecological conditions to which the experiment applies. The causal leverage attributed to human experiments in this situation is pointless, since the conditions of their ecological expression cannot be known with any confidence or precision.

The magic in arousal theory

Earlier we discussed the theoretical mechanisms attributed to social learning explanations. We return to that issue here. In *Seductions of Crime,* Jack Katz (1988) argues that to explain the attractions of deviance you must believe in magic. *Seductions of Crime* is a phenomenological approach to deviant behavior that emphasizes the foreground of experience in the commission of crimes. Where the majority of criminologists stress the causal role of such background factors as family conflict, community disorganization, and class conflict, Katz argues that the clue to explaining crime is the "sensory attractions of doing evil," the physical pleasures of sneaky thrills, righteous slaughter, and doing "stickup." Katz refers to "magic" as the process in which people let themselves be seduced by the criminal project. "To believe that a person can suddenly feel propelled to crime without any independently verifiable change in his background, it seems we must almost believe in magic" (1988:4). As a phenomenologist, Katz rejects causal explanations, but he stresses that people sometimes act as though they are forced by circumstances to behave badly. The key to the explanation is the powerful effects of emotions. In Katz's perspective, people sometimes let themselves be seduced by their emotions. They surrender to circumstances and situations, although he characterizes this as "an artifice" in the sense that this process involves an element of self-delusion. Persons are capable of resisting temptation but "give in" as though they are compelled by emotions. For example, a person who encounters humiliation can refashion it as an act of self-defensive rage, and strike out at the provocateur in "righteous slaughter." Intense emotional arousal deriving from the situation is at the core of the behavior. The impulse to kill is probably common in many of our social encounters but it is successfully stifled by self-control in most cases.

When we discussed social learning theory, I suggested that there may be a non-obvious process that operates in the cases of media exposure that is neither equivalent to classical conditioning effects nor operant conditioning. I would now add that it does not involve mimicry at all, but appears to result from an artifact of emotional arousal. The typical evidence for media effects discovers aggressive outcomes only when subjects are highly aroused (i.e, male subjects are insulted by a female confederate) and only when the experimenter *requires* the angry subjects to administer shocks to their female aggressor (under the pretext of a learning exercise). Fisher and Grenier (1994) attempted to replicate some of the critical work on pornography effects from Donnerstein (1983) that purported to show how certain themes in pornography facilitated aggression against women. Donnerstein reported that films that combined both aggression and erotic elements boosted shock levels given by male subjects to female targets higher than films containing either aggressive or erotic stimuli alone. In their replication of this work, Fisher and Grenier gave the subjects the option of skipping the learning experiment (and forgoing the administration of shocks) and proceeding directly to the debriefing. The vast majority of the subjects, even if angered, opted to skip this opportunity for retaliation and proceed towards the debriefing. In other words, the link between the stimulus films and the aggressive outcome reported in earlier work was an artifact of the design, that is, not even skin deep (see Fisher and Barak 1991).

But the nature of the aggression is far from clear. In an earlier work, Donnerstein reported that "aggressive behavior in subjects who have previously been angered has been shown to be increased by exposure to arousing sources such as aggressive–erotic films, physical exercise, drugs, and noise" (1980:279). Zillmann and Bryant (1982, 1984) used films of an eye operation and discovered that these boosted levels of aggression. Tannenbaum (1972:330) discovered that a humorous film had the same result. "Even viewing 'Sesame Street' or 'Mister Roger's Neighborhood' induced a threefold increase in aggression among preschoolers who initially measured low on aggressiveness" (Fowles 1999:28). However, the underlying process as understood by Zillmann and Bryant is called "excitation transfer." The subjects are angry because they have been insulted. The anger response tends to abate with time, but before it does so, it becomes reenergized by an intervening arouser. The pornographic films are important, not because they contain a message about women, as social learning theories suggest, but because they are highly arousing. Just as a stomach-churning film of an eye operation can transfer excitation to a previous source of arousal, so can an aversive noise, or a humorous film. The initial retaliatory impulses that arise from provocation are boosted by the intervening stimulus and the subjects react with more anger than they are probably aware of. But the effect is short-lived, since once the anger has abated, the aggression is no longer fueled by the emotional distress at the heart of the Buss paradigm. Also, in their longer-term (nine-week) study, Zillmann and Bryant suggested that pornographic films lost their ability to arouse (and fuel aggression) after repeated exposures, so that the ability of the films to promote retaliatory violence became naturally self-limiting. Writing about television, Zillmann came to the same conclusion. "It would thus appear likely that repeated exposure to dramatic portrayals of violent crime reduces rather than increases affective reactions" (1991:123).

The focus on arousal may be helpful in understanding one of Donnerstein's key findings. He reports that males respond more aggressively following aggressive–erotic exposure, but only when the target (i.e., the instigator) is female, as opposed to male. The inference that social learning theorists draw is that the subjects equate the females in the film with the female confederate, but it is just as plausible that the female confederates are a *greater* source of arousal *per se* than male confederates. And/or that males react more powerfully to insults from a woman than another man. Either way, the mechanism is the type and level of arousal, not social learning. As for the cognitive effects of viewing the pornographic films, we return to the Fisher and Grenier (1994) study, which sheds some light on this. This study was based on the same video that Donnerstein had used when testing for the effects of positive-outcome versus negative-outcome rape scenarios. Fisher and Grenier measured whether the various types of films (positive- and negative-outcome rape, erotica, neutral) were perceived differently in terms of the woman's apparent willingness to participate in the sexual activities and her apparent enjoyment. Perceptions varied significantly across the different film conditions as expected, but the expected changes in attitudes and fantasies were not found. Even though earlier studies suggested that extremely brief exposure to violent pornography causes men to fantasize about rape, and to increase acceptance of rape myths, there was no evidence of such outcomes in this study. Fisher and Grenier measured the aggressive/violent content in self-generated fantasies, violent content in Thematic Apperception Test scores, scores in attitudes toward women, acceptance of interpersonal violence, and rape myth acceptance. There was no difference across any of the treatment groups. Fisher and Grenier concluded:

> The current review on the literature on violent pornography, together with the current failures to observe effects of violent pornography on men's attitudes, fantasies, and behavior toward women, raises serious questions about the reliability of effects of violent pornography within the experimental procedures that have been used in research in this area.
>
> (1994:36)

And, just as the field studies of television violence show little consistent relationship to the acquisition of violent behaviors, the survey research on the association between violent pornography and antiwoman aggression has rarely indicated a link (Fisher and Grenier 1994:25; see also Scott and Schwalm 1988; Kutchinsky 1991; Diamond and Uchiyama 1999).

The media effects research, public policy, and law

The media effects research community has long lobbied for public policies to abate the harmful consequences of sex and violence in the popular culture, in comics, on televison, in pornography, and, most recently, in video games. Beginning around 1970, there was a series of national commissions of inquiry, primarily in the United States, also in the UK and Canada, devoted to researching and summarizing the effects of television violence on the one hand, and pornography on the other. But there was a remarkable lack of consensus about whether there were significant consequences to viewers from exposure to any of this material. On the television side, there was the 1968 National Commission on the Causes and Prevention of Violence. Verdict: television probably incites aggression, but no new research was commissioned. This was followed in 1970 by the President's Commission on Obscenity and Pornography. After two million dollars of new research, the verdict: pornography got a clean bill of health. Then the Surgeon General's Study of TV (1972) appeared with another million dollars of research. Verdict: television causes aggression in everyday life. When the National Science Foundation reviewed the conclusions, they replaced "TV causes aggression" with "TV *may* cause aggression in small numbers of individuals vulnerable to its influence." In 1979, in the UK, the Williams Committee into Obscenity examined the behavioral consequences of exposure to pornography – another clean bill of health. In 1982, the National Institute of Mental Health revised the earlier Surgeon General's report, and reported "a convergence of evidence" suggesting a causal role for violent television. In 1985, Canada's Fraser Committee of Inquiry into Pornography and Prostitution rejected the evidence of harm. The committee reported that "the research [on the effects of pornography] is so inadequate and chaotic that no consistent body of information has been established" – but a year later – 1986 – in the United States, Attorney General Meese's committee came to totally different conclusions. Why the differences? What appears to have made the difference was the increasing importance attached to the experimental studies and the decline in the use of field studies and criminological evidence linking media exposure to actual deviant outcomes. The experimental literature was to have a worrisome impact on the development of law.

In 1983 and 1984, Minneapolis and Indianapolis passed municipal ordinances to create liabilities for persons selling pornography that depicted "the graphic sexually explicit subordination of women, whether in pictures or words" or "in positions of servility or submission or display" (cited in De Grazia 1992:614). As Richard A. Posner (1988) noted, the law was much broader in its reach than the *Miller* test.[2] "The Bible contains many instances of what by contemporary standards is misogyny; so do *Paradise Lost* and *The Taming of the Shrew*, not to mention *Eumenides* – the list is endless" (De Grazia 1992:615). The ordinances were drafted by Catherine MacKinnon (1985) and Andrea Dworkin (1985), and supported by a coalition of conservatives, the religious Right, and radical feminists. The logic underlying the new legal approach came from the experimental effects literature, the literature of the Buss paradigm. It was the same literature that carried the day in the Meese Commission in 1986. However, given the robust protection of the First Amendment, the municipal ordinances went nowhere. The first was vetoed by the mayor of Minneapolis, the second was found unconstitutional by the U.S. Supreme Court in 1986. It is interesting that when the Meese Commission became an obvious pawn of the religious Right, the media experts in psychology tried to divorce themselves from the "overinterpretation" of the effects literature (see Linz, Penrod, and Donnerstein 1987). In other words, those who supported MacKinnon and Dworkin for politically correct reasons at the beginning of this policy crusade by basing their models on a concern for female victimization, divorced themselves from the crusade for politically correct reasons as the politics of censorship shifted to the Right (see Russell 1993).

The protection accorded speech in the United States provides some protection against such acts of censorship, at least in principle. As recently as 1990, Dennis Barrie, the curator of Cincinnati's Contemporary Art's Center, was charged under Ohio obscenity law for exhibiting a collection of Robert Mapplethorpe photographs, and 2 Live Crew was charged in Florida for obscene song lyrics (De Grazia 1992:654–56). Both cases led to acquittals, where charges should probably never have been laid in the first place. What the United States and Canadian cases illustrate is the power of interest groups to use the criminal law to advance their own values and interests. What we have not discussed is the role of psychologists and other media experts in assuming a role in this process.

Hidden agendas in scientific and moral leadership

Jib Fowles (1999) makes a convincing case for the idea that the academic industry devoted to the identification of negative media effects has an unacknowledged cultural foundation. Claims of dire consequences can be advanced on the weakest, most inconsistent evidence because the condemnation of violence is based, not on science, but on what we called in earlier chapters "extra-scientific" incentives. Media violence is, in Fowles's term, "a perfect whipping boy" because media changes are such a large target (even if we need to exaggerate the level of violence), because hostilities in fact or fiction are provocative, and because "the issue attracts no supporters. Virtually no one speaks out in defense of television violence … as a whipping boy, television violence could hardly be improved" (1999:55). Martin Barker (1984, 1989) made a similar argument for the cultural foundation of the anti-comics campaign in the UK in the 1950s. This was led by the Comics Campaign Council and the National Union of Teachers. It focused less on crime comics (which were said to be criminogenic in the US) than on horror comics, and was an expression of the rejection of American cultural invasion of the UK and the imperilment of its youth. Entertainment based on television violence and horror comics represent the culture of the plebeians, society's cruder elements, the great unwashed face of unruly youth. Fowles points out that real class antagonisms have diminished tremendously in the past century, without disappearing entirely. Borrowing from Bourdieu, Fowles argues that the preoccupation with material comfort has been overcome by a new kind of capital – cultural capital, the sense that what elevates status in contemporary society is the acquisition of refinement, an ability to make distinctions based on taste and a heightened moral sensibility. A condemnation of physical or sexual violence, especially based on expertise regarding its effects, socially elevates those who make it, and reinforces their cultural capital.

> Television violence [is] an issue in the largest social struggle – that of the privileged (the baccalaureates, the dominant) against the rest (the dominated). Television violence is the rhetorical issue of choice in the dominants' efforts to demean and control the dominated.
>
> (Fowles 1999:58)

But the dominant and the dominated are not real classes – they are postures created by defining the social good effectively. The academy creates a social momentum for its visions through its ability to legitimize knowledge and define the good, and set an authoritative perspective for the rest of society.

The prestige of the academy is further enhanced by setting the pace for normative controls in the legal order, redefining the kinds of images that harm, the forms of control that are "justified," and making the social development of children and women dependent on a scientifically grounded agenda. Besides law, the prestige of the academy grows further by creating alliances with elites in other influential professions, such as medicine. This explains the academy's interest in having its views certified by the Surgeon General and such institutions as the American Medical Association, thereby expanding a psychological question of media effects into a question of "public health" (Mulvey and

Haugaard 1986). Recall that the anti-comic crusade mentioned earlier was led by Dr. Frederic Wertham, chief of psychiatry at New York Hospitals in the 1940s.

The latest "threats" to public safety

In assessing the challenges to psychology in the 21st century, we would be remiss if we did not acknowledge dramatic changes in the mass media that have significantly impacted the lives of ordinary people. The first is the normalization of pornography. Adult entertainment sites are among the busiest web pages in the world. People may go to Google more often, but Google is a universal search engine. Porn Hub (2019) is rather more focused. In 2019 there were 42 billion visits to the site – over 115 million per day. However, what seems to have caused the greatest alarm in the study of media effects has been the rise of violent video games over the past two decades. Digital media have made available an increasingly array of violent games that have attracted a large youthful following. Players get to simulate shooting other armed combatants, zombies, space invaders, and to stalk and murder predators, to spill blood and explode body parts graphically on the screen, to chop the victims into pieces and violate them sexually. Many games have been described as "murder simulators" which train young people to become sadistic murderers. Critics of violent videos stress that the impact of violent video games is probably larger than simple exposure to violence in entertainment in other media, such as television or films, since the person is not passively observing but is actively engaged in the action through the joysticks that control aggression. In addition, the games can be played on a variety of devices including computers, tablets, consoles, and smartphones.

One of the critical public discussions on the link between violent video games started in the aftermath of the 1999 Columbine School massacre in Colorado, where two young men, Eric Harris and Dylan Klebold, murdered twelve fellow students and a teacher with semi-automatic weapons before taking their own lives in the school cafeteria. It was widely reported that the young men were immersed in violent videos, particularly the games, *Doom* and *Quake*, and were making their own violent videos. A subsequent FBI investigation suggested that the perpetrators had serious mental problems: Harris was psychopathic, and Klebold was a depressive. Nonetheless, the link to violent video exposure persisted. In the aftermath of the Parkland School shooting in 2018, President Trump blamed the violent media. The shooter in the Parkland case, Nikolas Cruz, who murdered seventeen persons at the school, reportedly was obsessively attached to violent video games which he played for hours every day (Schipani 2018). He had also been diagnosed with, and was being treated for, mental health disorders.

The theory proposed for the media–violence link which updates social learning is called the General Aggression Model. It holds that violent media produce short-term increases in physiological arousal, enhances feelings of aggression and "primes" the person at a sub-conscious level to mimic the aggression. The theory has been enthusiastically endorsed by the American Academy of Paediatricians and the American Psychological Association media watchdog organizations. The theory is at dramatic variation with perspectives in criminology. Writing in *Criminal Justice and Behavior*, Savage and Yancey (2008) report a meta-analysis of the media–crime link and concluded that media studies that controlled for "trait", that is, variations in individual disposition to aggression, did not report changes arising from exposure to media aggression.

There have been many attempts to pull together all the different kinds of research on violent media and subsequent outcomes. One of the most impressive was prepared by Craig Andersen and his associates (2010), looking at these linkages in both Japanese and Western societies. Their study identified 381 effect-size estimates based on 130,296 participants. They examined experimental, longitudinal, and correlational studies separately, and distinguished research that followed "best practices" from those with less than optimum practices. They focused on increased aggressive behavior, aggressive cognition, and aggressive affect and for decreased empathy and prosocial behavior.

As for their main findings, Andersen et al. (2010:167) concluded that video game violence was positively associated with aggressive behavior, aggressive cognition, and aggressive affect, and negatively associated with prosocial acts and empathy. In terms of effect sizes, Andersen et al. suggest that the relationship between violent video game exposure and aggression is $r = +0.152$. That is a bivariate estimate that does not always have much relevance for understanding behavior in everyday life since it fails to control for the fact that the games are much more popular with males, and males are already more predisposed to direct aggression than females. The "coefficient of determination" – the R^2 – is 2.3% explained variance. In an equation that includes the other main predictors of violence, this factor becomes vanishingly small.

The first important academic response to the meta-analysis was published by Ferguson and Kilburn (2010) in an article titled "Much ado about nothing." Ferguson and Kilborn accept the finding of $r = +0.15$ but point out that the Andersen meta-analysis failed on a number of other measures. The inclusion of experiments where the "aggression" is trivial leads to an overestimation of the effect size. They suggest that when one examines studies with more realistic behavioral measures of violence, the bi-variate correlation is more like $r = +0.04$. There is also a major problem with a selection bias towards inclusion of results that show a positive correlation – and which are, for that reason, more likely to end up being submitted to journals and entering the received literature. This extends to the use of unpublished results which are selected through contacts with authors who are already invested in establishing a causal linkage. Andersen did not contact Ferguson to obtain his unpublished *negative* results. In addition, when one examines the correlation over time between the growth of violent video games from 1998 to 2007 (based on sales) and the youth crime for violent youth in the US, the correlation is $r = –0.95$ – an almost perfect inverse relationship!

Brown v Entertainment Merchants Association (2011)

In 2005 the State of California passed a law to prohibit the sale or rental of violent video games to minors. The legislation covered games that made a range of options to the player or players, which included "killing, maiming, dismembering, or sexually assaulting an image of a human being" where a reasonable person would infer that this appealed to the deviant nature of a young person or was patently offensive to community standards. The California legislators were motivated to regulate the violent video games because of their alleged association with deleterious consequences for youth. The matter was decided by the Supreme Court of the United States in 2011 in *Brown v EMA.* There are various constitutional reasons for protecting free speech, even if it is unsavoury and offensive. If there is an "actual problem" in need of solution, a challenge to free speech could be heard by the court. But "California cannot meet that standard" (Brown 2011:12). It cannot show a direct causal link "between violent video games and harm to minors" (2011:12). California's attempts to show that link relied on the research of Dr. Craig Andersen, whose research attempted to link exposure to violent video games and harmful effects on children. "These studies have been rejected by every court to consider them, and with good reason. They do not prove that violent video games *cause* minors to *act* aggressively." They report correlations between violent exposure which are "miniscule" – like making louder noises a few minutes after media exposure. These experimental effects are small and comparable to those found in other media. "Dr. Andersen admitted that the 'effect sizes' of children's exposure to violent video games are 'about the same' as their exposure to violence of television" (2011:13). He further admitted that the same effects have been found for children exposed to Bugs Bunny and Road Runner cartoons, and after exposure to video games designed for young children. This is contrast to the oft-stated opinion that the link between violent video games and subsequent aggression is as strong as the link between cigarette smoking and lung cancer.

The Brown decision appears to resonate more with the perspective of Ferguson and the skeptics of media effects. Ferguson (2016) rejects the General Aggression Model, which purportedly acts like a hypodermic needle to inject attitudes, feelings, and reactions automatically and subconsciously. He proposes a "Self Determination Theory and Mood Management Theory," which recognizes that viewers select media to meet their expectations and goals, typically observing the age-graded classifications that advise viewers of media content. Ferguson wrote that, in 2013, "a group of 238 scholars asked the APA to retire its various policy statements on media violence, because of the mismatch between these statements and the available, often conflicting statements." In 2015, the APA (2015b) seemed to meet them half-way by claiming both that their review "confirms [the] link between playing violent video games and aggression" while also acknowledging that it "finds insufficient research to link video game play to criminal violence." The most recent meta-analysis by Hilgard, Engelhard and Rouder (2017) re-examined the original Andersen analysis and concluded that the evidence for short term effects of violent games on affect and behavior was "overstated." This was due to publication bias, p-hacking of the results to maximize the likelihood of statistical significance and selection bias in the decisions on which studies to include and exclude. After making adjustment for such biases, they wrote that "the effects of violent video games are likely smaller than anticipated and may be so small (r = 0.02–0.15) as to be very challenging to detect in most experiments" (2017:769). They also note that studies which hope to capture the important mediating and moderating effect of the media outcomes (the things that need to be controlled before linking media and violence) would require, on average, hundreds more subjects than employed by the crop of existing studies (2017:770).

Notes

[1] Childhood in such a city must have been a Roman psychologist's nightmare.

[2] In the United States, the definition of obscenity was laid down in the U.S. Supreme Court in the 1973 case of *Miller v. California.* Miller held that to be obscene, the material must (a) appeal to a prurient [or obsessive] interest in sex; (b) contain "patently offensive depictions or descriptions of specific sexual conduct" as judged by a local grand jury in light of the contemporary standards in the community; and (c) when taken as the whole, have "no serious literary, artistic, political or scientific value." The scope attached to the third criterion has made convictions for obscenity difficult to sustain. In Canada, the definition of obscenity was laid down in the Supreme Court's decision in *Regina v. Butler* (S.C.R.) 1 432. Canadian law forbids "the undue exploitation of sex" – implying that exploitation that is "due" is legal. The determination of what is "due" is a question of national community standards of what individuals would tolerate their neighbors to see. The *Butler* case built on a series of lower court cases that allowed expert social science evidence about the harmful effects of pornography. The new decision suggested that materials that were harmful would not meet the community standard test. That would be the case even if such works had any serious literary or other value, as in the United States. U.-S. constitutional law places an extremely important role on protection of free speech that borders on treating it as an absolute good, while Canadian law is designed to balance competing interests (free speech and individual security).

8 Gender and psychology

From feminism to Darwinism

Introduction

In this chapter, we move away from a strict experimental psychology to examine issues of gender that have emerged in the classical period. We focus on two developments. The first is associated with the work of Carol Gilligan, and the suggestion that moral development is quite different in males and females, with the result that men and woman differ significantly in how they make moral choices. The second involves the evolutionary psychology of David Buss and others, which makes equally radical claims about how parental investment in human reproduction and sexual selection pressures have differently shaped the preferences and morals of human males and females. Both perspectives have generated a great deal of debate over the question of whether such differences actually exist, and, if they do, whether they are innate, socially acquired, or a conjoint outcome of each. Both perspectives suggest that they may provide the basis for fundamental changes to society. Gilligan writes:

> The rash of questions about relationship and difference which become inescapable once women enter the conversation are now *the most urgently pressing questions* on the local, national and international scene. The political has become psychological in the sense that men's disconnection and women's dissociation perpetuate the prevailing social order.
>
> ([1982] 1993:xxvii, emphasis added)

As women recover their distinctive voice, their resistance argues for, in Gilligan's words, "potentially revolutionary" change, and threatens the demise of patriarchal societies that are based on men's disconnection from women and women's dissociation from their own distinctive moral voice, a voice based on their sense of connectivity and care for others.

On the side of evolutionary psychology, Buss writes that

> we are empowered now, perhaps more than at any previous time in evolutionary history, to shape our future … We are the first species in the known history of three and a half billion years of life on earth with the capacity to

> control our own destiny. The prospect of designing our destiny remains excellent to the degree that we comprehend our evolutionary past. … Only by understanding why these human strategies have evolved can we control where we are going.
>
> (1994b:220, 222)

Although lacking the specific initiatives associated with feminist psychologies, Buss's claim is no less grandiose in its scope. We begin our work with an examination of the historic debate over the crisis in development in young males and females in modern society.

Gilligan, Sommers, and *The War Against Boys*

Throughout the 1990s, the American Association of University Women (AAUW), with assistance from the Ms Foundation and other women's interest groups, mobilized a media-savvy campaign to communicate the message that female students were being seriously undermined by the nation's educational institutions. The AAUW published a study, *How Schools Shortchange Girls* (1992) that suggested girls were being disadvantaged in teaching processes, and that this was resulting in a diminished sense of self-esteem as adolescent girls reached adulthood. Advocacy experts suggested that teachers were biased in favor of males in the classroom, and that boys were permitted to "cry out" responses in class, "eight times" more often than girls, according to the Sadkers (the researchers) (AAUW 1992:68), while girls who did likewise were told to raise their hands if they wanted to speak. Clinical psychologists reported that there was a dramatic shift in adolescent female suicide, suggesting that the popular culture was "girl-destroying." The image conveyed by the campaign was that little girls showed self-confidence, insight, and sparkle in their eyes in their formative years, but faced a downward spiral in self-worth as they entered adolescence. The exuberant girl of primary school became the shrinking girl of high school.

The media took to the release of such provocative "information" with enthusiasm. The story of the shrinking girl was reported uncritically in many of the leading newspapers and magazines. The initial study that cost $100,000 was publicized by AAUW with a budget of $150,000. The U.S. Congress passed the Gender Equity in Education Act in 1994. "Millions of dollars in grants were awarded to study the plight of girls and learn how to cope with the insidious bias against them" (Sommers 2000a:23). There was a backlash against boys, since a subtext of the campaign was that the psychological deficits faced by girls were a result of the advantages conferred unfairly on boys. These advantages were presumably one of the devices that guaranteed "the reproduction of patriarchy." Gradually, part of the pedagogical agenda to restore the equal treatment of girls was to reconstruct boyhood, to render boys "less competitive, more emotionally expressive, more nurturing – more, in short, like girls" (2000a:44). Ironically, those who argued that boys were being advantaged in the socialization process subsequently argued that the boys were being poisoned by their own masculinity, since the latter was equated not merely with competition, but with violence, indifference to others, and lack of connectivity to women. The impression created by this line of thinking was that gender is essentially a caste system with two qualitatively different kinds of human beings, males and females. The system is grounded in different patterns of psychological development that provides the lynchpin that confers systematic advantage to the upper caste, males, and that stifles the growth of the lower caste, females. This analysis implies that masculine and feminine identities are no more than social constructs that can be changed by policy, although, as we shall see, there is ambiguity on all sides in regard to this perception.

The evidence of the marginalization of girls in American education was first criticized at length in Christine Hoff Sommers's *Who Stole Feminism* (1994) and later in Judith Kleinfeld's *The Myth That Schools Shortchange Girls* (1998). In a spirited attack on the empirical evidence in the 1992 AAUW report, Kleinfeld argued that

> the findings in this report are based on a selective review of the research.
>
> Findings contrary to the report's message were repressed. These contrary findings indeed appear in studies the AAUW itself commissioned, but the AAUW not only did not include these findings in their media kits but made the data difficult to obtain … Major assertions in the AAUW report are based on research by David and Myra Sadker that has mysteriously disappeared. Evidence which contradicts their thesis that the schools shortchange girls is buried in supplemental tables obtainable only at great difficulty and expense. Such shady practices undermine public confidence in social science research.
>
> (Kleinfeld 1998:2, 6)

Expanding on her earlier research, Christina Hoff Sommers in *The War Against Boys* (2000a) argued at length that the evidence for the educational deficits faced by girls was contradicted by the facts:

> Data from the U.S. Department of Education and from several recent university studies show that far from being shy and demoralized, today's girls outshine boys. Girls get better grades. They have higher educational aspirations. They follow a more rigorous academic program.
>
> (ibid:24)

Sommers reported that girls read more books, showed higher levels of artistic and musical ability, were more likely to study abroad and join the Peace Corps. By contrast, boys were far more likely to leave school prematurely, to receive discipline at school for misconduct, and to receive "special education" for learning deficits and to show signs of hyperactivity and attention deficit disorders. The National Center for Education Statistics (NCES) reported, as mentioned in Sommers (1994), that boys were more likely to go to school unprepared, that is, arrive at school without books or paper and pencils, and to appear without completing homework. The NCES reported in 1996 that girls were more likely to devote longer periods of time to homework, and to do so at every level of schooling. This is reflected in the long-term trends. In 1976–77, slightly more males had earned baccalaureate degrees than females: 494,424 versus 423,476. However, by 2016–17, the number of males earning such degrees had increased by 169% (836,045) while the number of females earning such degrees had increased by 264% (1,119,987) (NCES 2019: Table 322.20, p. 336). If there was a gender disparity in educational achievement, it appeared to favor females. This was also evident at the Masters and the Doctoral levels. From 1976–1977 to 2016–2017, the number of Masters degrees awarded to men increased by 189%, while they increased by 319% for females. Over the same period, the number of Doctoral degrees earned by males increased by 118%, while it increased by 495% for females (NCES 2019: Tables 323.20, p. 339 and 324.20, p. 342). At all three levels, the *rates* of growth in educational attainment, as well as changes in the *absolute numbers* of degrees, favored females. The same trends are evident in the UK (HEPI 2016) and Canada (CBC 2011; StatsCan 2016).

In 1998, the National Council for Research on Women in the US published *The Girls Report: What We Know and Need to Know about Growing Up Female.* The report dismissed the problem of self-esteem differences across gender, bringing into question the very utility of the concept. The report failed to replicate earlier studies of females' diminished self-esteem in adolescence (Kleinfeld 1999:18). The report was based, in part, on the research of University of Denver psychologist, Susan Harter, who studied 900 male and female students in grades six through twelve. She found no evidence for "loss of voice" for female adolescents, or any evidence for gender differences favoring females.

One of the more worrisome pieces of evidence that Kleinfeld and Sommers presented was a 1990 survey of gender roles and self-esteem conducted by the AAUW. This survey provided evidence that, in the views of the girls and boys themselves, girls were systematically favored by their teachers. Both girls and boys believed that teachers thought girls were smarter than boys, were more likely to be complimented by teachers, less likely to be disciplined, more likely to be called on in class, more likely to get the teachers' attention, and preferable in terms of whom the teachers liked to be around. In other words, while lobbying the public on the image of the "short-changed girl" drowning in a sea of sexism and facing educational deficits at every turn, the AAUW had in its possession survey information that suggested quite the contrary. Not only were girls outperforming boys in terms of academic achievements, they were experiencing greater levels of self-esteem, as indicated by the survey data.

What about the patterns in adolescent suicide? In 1997, there were approximately 4,500 suicidal deaths in the United States; 84% were males. In fact the long-term trends in suicide in the US from 1950 to 2017 (Elflein 2019), in Canada from 1950 to 2009 (StatsCan 2012), and in the UK and Australia (Schumacher 2019) indicate that the male suicide rate has been typically 3–4 times higher than females. As for the Sadkers' "evidence" that boys were permitted to "call out" eight times more often than females, this claim was based on an unpublished paper that never made it to the refereed literature. The original research commissioned for the National Institute of Education either could not be found or failed to corroborate the claim of gender bias in "call-outs" (Sommers 1994:164ff.). The claim of the "short-changed girl" was, in Sommers's words, "politics dressed up as science."

What is the harm of this bias? Surely, advocacy on behalf of females by women's organizations can only do good. Kleinfeld argues that the campaigns of the AAUW and similar groups actually insult women by understating their accomplishments. On the positive side, they identified the lag of female accomplishments in science and mathematics, but this gap is narrowing, while the gap between male and female language and composition is not. "Unfortunately, the feminist agenda, because it is pushed so strongly and receives so much attention from media elites, distracts us from the real problem of low educational achievement among African-American males and boys more generally" (Kleinfeld 1999:19). That remains true today.

Sommers argued that the ideological foundation for the "shrinking-girl" thesis was found in the psychological research of Harvard psychologist, Carol Gilligan. Gilligan is a student of Lawrence Kohlberg, a leading proponent of the theory of moral development. Kohlberg argued, following Piaget, that as children grow older, they not only show signs of more complex levels of cognitive skills, but become increasingly sensitive in terms of *moral* development. Gilligan discovered what she claimed to be gender differences in patterns of moral development. Unlike most of the theories examined in social psychology, this perspective had roots that grounded developmental patterns, not in behaviorism or cognitive theory, but in psychoanalysis.

Kohlberg's moral development theory

Kohlberg originally thought he could identify six discrete levels of moral development, although recent work suggests that there are only four major levels that reflect clear developmental progression (Greeno and Maccoby 1986:311). As with Piaget, such levels were thought to be temporally sequential and hierarchical. Level three, which marks a movement to adult reasoning, is characterized by a preoccupation with bonds with others and the development of trusting relationships. This is the level at which females were alleged to "top out", that is, caring for others. Level four reflects more societal concerns for the rule of justice and law, and maintenance of the collective interests of society. Level four, for Kohlberg, reflected a more global sensibility that transcended obligations and attachments idiosyncratic to individuals and their personal ties and marked a level of growth limited largely to males. As a student of Kohlberg, Gilligan departed from the master by identifying the coding system which relegated women to a lower level of moral development as "androcentric," and by postulating that women progress on a path of moral reasoning different from men. The differential experiences and obligations of women as caregivers in contemporary society heighten the importance of connectivity for them in a way that overshadows the more abstract male concerns. "This different construction of the moral problem by women may be seen as the critical reason for their failure to develop within the constraints of Kohlberg's system" (Gilligan [1982] 1993:19). The key mechanism for gender differentiation was not genetic predisposition, but patterns of identity formation in girls and boys raised by their mothers. According to this theory, girls more closely identify with mothers than do their brothers, and presumably experience greater interconnectedness with them. Boys, being more aware of their separate identity, differentiate themselves more completely from their mothers and experience generational power imbalances that valorize the importance of justice and equality as opposed to an ethical sense based on attachment and care (see Chodorow 1978). In the result, female children bond and connect while male children individuate and become focused on rights.

Gilligan's research was based on a handful of small-scale studies: the college student study designed to explore identity and moral development in early adulthood (*n* = 25), the abortion decision study designed to explore the reasoning of young women facing unplanned pregnancies (*n* = 29), and the rights and responsibilities study designed to explore moral conflicts, individual choices, and judgments of hypothetical moral dilemmas at different ages over the life cycle (*n* = 144). The results of her work are reported in a discursive manner, citing suggestive quotations from the subjects, but there are never any explicit hypotheses, clear design features appropriate to testing them, or tests of statistical significance to determine whether measurable differences exist.

The social science literature contains many studies of gender differences on such traits as empathy and altruism (Hoffman 1977; Eisenberg and Lennon 1983). In addition, there are striking differences in social relationships in male and female peer groups and how they engage in play (Maccoby 1985). There are also tremendous differences in childhood aggression in males and females at every age. However, the evidence of differences in moral reasoning fails to substantiate Gilligan's claim of a distinctive moral voice. Lawrence Walker (1984) reviewed sixty-one studies that tested for gender differences of the sort Gilligan suggested. They failed to establish that males scored higher than females on Kohlberg-type scores.

> In adulthood, the large majority of comparisons reveal no sex differences. In the studies that do show sex differences, the women were less educated than the men, and it appears that education, not gender, accounts for women's seemingly lesser maturity … There is no indication whatever that the two sexes take different developmental paths.
>
> (Greeno and Maccoby 1986:312)

In a review of Gilligan's book in the *Merrill-Palmer Quarterly*, Colby and Damon (1983) came to similar conclusions. "There is very little support in the psychological literature for the notion that girls are aware of others feelings or are more altruistic than boys" (1983:475). Colby and Damon recount the differences in children's play. Boys like games with organized rules and competition. Girls like "dyadic intimate exchange and turn-taking games." But it is not clear why. The organized games may lend themselves to dominance displays and bullying, but it is impossible in naturalistic observations to determine whether such differences are a natural gender trait or whether these are the molds into which the children are shoehorned by parents and schools. There are also many other differences in such areas as occupational choices, aggressiveness, and competitiveness, but these may reflect little more than opportunities and restraints, and they are differences that are increasingly being narrowed in contemporary society.

What about Gilligan's supposition that Kohlberg's system artificially privileges "justice" thinking over "relationship" thinking? This would seem to be a credible concern if, in fact, the empirical evidence corroborated such gender differences, but the more systematic tests of gender differences failed to corroborate the claim. In the alternative, Gilligan treated the readers to reports from her handful of qualitative studies. However, as Colby and Damon point out, such reports were based exclusively on *anecdotes* that appear to have been chosen to illustrate the differences on which the research was premised. *In a Different Voice* contains no information on how the respondents' views were coded to determine whether the "evidence" was selectively cited to confirm Gilligan's views. The abortion study was particularly problematic, since it is unclear that a small sample of young single women (n = 29) facing unplanned pregnancies and contacted via a counseling service designed to deal with the potential stress or trauma of confronting abortion – sometimes for a second and third time (and sometimes with the same married man) – is a valid source of general gender differences in moral reasoning. In particular, a specific gender comparison is, at one level, out of the question, since men cannot have abortions. But the study also ignored information from the *potential* fathers, whose views on aborting potential offspring might have provided a measure of difference in moral reasoning. It also overlooks the potential bias that arises from sampling exclusively in an abortion counseling clinic. The absence of comparative data jeopardizes the claim to gender differences of all but the most trivial kind. "Although Gilligan's abortion interviews yield some interesting data on real-life decision-making processes, they do not provide support for her thesis of sex bias in Kohlberg's theory" (Colby and Damon 1983:478). Colby and Damon conclude by warning about the irony of Gilligan's position. The idea that men and women reason in qualitatively different ways may tend to justify gender stratification by relegating social differences in opportunities and achievement to innate differences: it is "important to guard against reinforcing gender stereotypes that in themselves contribute to the maintenance of women's oppression" (1983:480). Sommers (2000a) went further. After repeated attempts to obtain copies of the original research protocols and raw data, she concluded that the data did not exist.

Gilligan herself qualified the relevance of the abortion study. Since the focus of the study was the relationship between judgment and action,

> … no effort was made to select a sample that would be representative of women considering, seeking or having abortions. Thus the findings pertain to the different ways in which women think about dilemmas in their lives rather than to the ways in which women in general think about the abortion issue.
>
> ([1982] 1993:82)

Gilligan seems to think that this ameliorates the methodological limitations of her research design. But it does not. Her position amounts to the claim that abortion decisions can be used to throw light on women's *general* moral dilemmas. On the one side, many women refuse to consider abortion at all under such circumstances (see Luker 1984). And, on the other, in her study, it is not clear that abortion automatically presented a traumatic dilemma for the women in counseling.

Furthermore, if no care was taken to grasp how women thought about a particular dilemma, abortion, on the strength of what would we be permitted to make valid inferences about the larger topic of the "different ways" in which women think about dilemmas in general? Gilligan seems to imply that shoddy research has more power when it comes to larger questions. Luria, writing in *Signs*, thinks otherwise: "In general, Gilligan's sample specification is inadequate to justify her group characterizations" (1986:317). In other words, women's thinking about abortion, particularly contacted in this context, cannot be used to gauge important gender differences in moral thinking or moral dilemmas. Gilligan's analysis also failed to shed light on the *different* responses among the women, some of whom chose abortion while others did not. Her analysis fails to explain the determinants of choosing abortion versus alternative courses of action. A constant, "female connectivity," cannot explain a variable.

Gilligan's work appeared with Harvard University Press in 1982. It was a tremendous academic *cause célèbre*, but it also spawned a torrent of empirical criticism from many quarters in the immediate years after its release, including criticisms from feminist psychologists and gender theorists sympathetic to women's political and economic advancement. It was reissued in 1993 without a hint that anything important had occurred in the discipline in the intervening years to raise questions about Gilligan's theory. In 1986, Thoma had written: "There is now considerable evidence that justice defined measures of moral reasoning are not biased against females. Further, there is little support for the notion that males are better able to reason about hypothetical dilemmas [than females]" (1986:176). Systematic reviews of the moral dilemma literature found no support for her claims about different moral voices (Brabeck 1983; Gibbs, Arnold, and Burkhart 1984; Friedman, Robinson, and Friedman 1987). Martha Mednick wrote: "There seems to be general agreement among moral development researchers that the presumed sex differences have not been supported" (1989:1119). The subtitle of her paper was instructive: "Stop the Bandwagon, I Want to Get Off." Luria came to similar conclusions: "When usual summary techniques are applied to add all the studies together, the data do not support any finding of a statistically significant sex difference" (1986:318). Gilligan reported in the "Letter to Readers" section in the 1993 edition that she did not revise the study "because it has become part of the process that it describes." Which was what? "The ongoing historical process of changing the voice of the world by bringing women's voice into the open, thus starting a new conversation." But were Mednick's, Maccoby's, Luria's, or any other of the women's voices even acknowledged? Not a word. The book became part of the process it described by indifference to criticism, not by conversation.

Gilligan's indifference to the empirical evidence brings us back to a theme that we have found throughout social psychology: the moral lesson frequently embedded in psychological research outweighs its empirical foundations. In our review of the classic group influence studies, from Sherif to Asch to Milgram to Zimbardo, we showed how the key studies at the heart of the tradition were *not* careful experiments in the model of the natural sciences. They were demonstrations, typically undertaken without the identification of specific hypotheses, and derived with little benefit from psychological theory. So, in that respect, the current work *demonstrates* the differences Gilligan attributes to her subjects through impressionistic storytelling. Just as the classic work had, for various reasons, a powerful moral appeal, similar processes are operative here. What are they?

Mednick's bandwagon hypothesis

Writing in 1989, Martha Mednick argued that the three most prominent bandwagons in psychology in the previous two decades surrounded issues of gender: women's fear of success, the emergence of androgyny as an alternative to typical gender constructions and "different voices" in the moral thinking of males and females. Mednick argues that the staying power of the bandwagon is "quite independent of scientific merit" (1989:1120). The concept of distinctive moral voices in men and women has tremendous appeal because it plays into familiar gender stereotypes, the belief that distinctive female perspectives are excluded by male-dominated sciences, that women's more sensitive moral compass is stifled by the harsh realities of patriarchy, and that important social change will only be possible when women assume political dominance. However, as Mednick notes, the stereotypes are stronger than the real gender differences. Mednick summed up as follows: "the simplicity of such ideas is appealing; such gender dichotomy confirms stereotypes and provides strong intuitive resonance" (1989:1122). The gender polarity that is presupposed by Gilligan's analysis has been superseded by attempts to transcend the presupposition of essential differences between women and men (Prentice and Miller 2006).[1] Male–female bivalence has been challenged by the diversity in sexual identities.

In 2000, Jaffee and Hyde published a meta-analysis of research on gender differences in moral orientation to assess quantitatively the evidence for care orientation among females and a justice orientation among males, as had been predicted by Gilligan's work. Their search identified 113 empirical investigations which yielded (a) 160 independent effect sizes in care orientation (based on 5,783 males and 6,654 females) and (b) ninety-five independent effect sizes for justice orientation (based on 3,831 males and 4,307 females). There were small differences in care orientation favoring females (*d* = -0.28) and justice orientation favoring males (*d* = 0.19) but the effects were small.[2] They concluded that the findings "do not offer strong support for the claim that the care orientation is used predominantly by women and that justice orientation is used primarily by men" (2000:703). The differences are also influenced by things such as age, socio-economic status, and how the measurements are actually taken. Walker (2006:109) later concluded that "gender explains a negligible amount of the variability in moral reasoning development. It is time to set this issue aside." Later investigators began to explore new sources of moral judgment and moral action, and developed related concepts such as "ethical sensitivity". Differences were identified in men and women, particularly in the context of professional socialization in services such as dentistry and teaching. Many professional groups developed scales to tap issues such as racial bias and other sources of potential insensitivity among recruits to the profession. You, Maeda, and Bedeau (2011) reported findings of their meta-analysis of gender differences based on results from these unstandardized professional sensitivity scales. They included nineteen studies (about 2,000 males and females in total) and found a small gender difference (*d* = 0.24). Ongley, Nola, and Maiti (2014) found small differences in "donation behaviors" in children aged four and eight – the older children and the girls were more generous in allocating "stickers" in a donation game.

The new studies have moved away from Gilligan's original approach to moral development, but they have not abandoned the investigation of important gender differences in important areas of life related to ethical issues. For example, women are significantly more likely to make real-world online donations to charities as measured by responses to online GoFundMe fundraisers. They are more empathetic than men (*d* = –0.27). And they scored lower on major negative personality measures such as Machiavellianism (*d* = 0.27), narcissism (*d* = 0.16) and psychopathy (*d* = 0.67) (see Schmitt 2019 for an overview of these areas). Many of these differences are moderated in part by context (i.e., age, nationality, status of women, etc.). There is also robust evidence of gender differences in aggression (Archer 2000) and in sexual behaviors (Petersen and Hyde 2010). However, these observations do not decide the question of why such differences occur. This takes us to one of the great dilemmas in contemporary psychology – the degree to which distinctive human characteristics are developmental stages, fixed traits, learned, or some combination of these and other factors.

Gender and the rise of evolutionary psychology

There are some ironic parallels between the feminism of Gilligan and the Darwinism of David Buss (1994, 1995) and the new generation of evolutionary psychologists. Both treat gender differences in sexual orientation as real, although the mechanisms underlying them are quite different. Unlike most psychology, neither work is experimental nor borrows primarily from experimental evidence. Both speak to fundamental questions of how sexuality structures other elements of social life. And evidence, the stock and trade of scientific life, in both cases, tends to raise more questions than it answers. Evolutionary psychology has been one of the most important new intellectual developments in contemporary psychology, and requires a close examination to evaluate its potential for the intellectual development of social psychology.

Evolutionary theory in biology and medicine is the foundation for intellectual growth in those areas. It is materialistic, non-teleological, and non-essentialist. Evolutionary models explain changes in species over time as a function of biological variability, often the result of chance variations in traits, and selection pressures that result in a greater or lesser reproductive success. Th*e fitness* of an organism is the degree to which the organism's inherited characteristics contribute to its reproductive success. Evolution is non-teleological in the sense that adaptive changes are not directed *a priori* toward some state of perfection or transcendence. The theory is non-essentialist inasmuch as it denies that there are specific rigid traits that define a species. The evidence suggests instead that there is variability in traits within a species, greater variability between related species even though related species share many of the same genes, and that the form of the organism is constantly evolving under changing environmental pressures.

Arguments of this sort in respect of the evolution of the *physical* characteristics of species such as teeth, bones, muscles, organs such as eyes, stomachs, fingers, and toes are universally accepted in contemporary science. And arguments to explain the *social* behaviors of insects and animals are standard in the curriculum of biology, entomology, and zoology. The attempt to explain *social* characteristics (altruism, the sense of justice, jealousy), particularly in the human species, has often met with resistance, particularly in the social sciences (Gould 1978; Rose and Rose 2000). It is assumed that the basic adaptive mechanisms in human experience derive from two major sources: Pavlovian classical conditioning and Skinnerian operant conditioning. It is often assumed that the understanding of human social life requires no knowledge of human evolution, as though human social evolution stopped with the natural selection of the capacity to learn. This assumption often co-appears with the related dubious assumption that human traits are either biological (genetically determined) or environmental (the results of socialization), either nature or nurture, as opposed to a complex interaction of both mediated by history and culture (Dupré 2003).

Evolutionary psychology is premised on the idea that human *social* behavior has evolved under natural selection pressures, and that human conduct has large instinctual foundations or elements. The approach to analyzing the mechanisms that influence social behavior is the same sort of "reverse engineering" that applies to the analysis of the physiological properties of the organism. We examine, for example, the teeth of a species as well as the realities of the food supplies. Where the ecology supports savannahs and prairies of vast grasslands, molars for grinding grains, seeds, and other plant foods show "design features" that exploit the food resources efficiently, as in elephants and bison. Over millennia of variations in tooth morphology, those animals best able to exploit the resources would tend to enjoy higher levels of fitness through a process of natural selection, since such morphological traits are preserved across generations in the species' genes. By contrast, animals with highly developed canines evolve in environments that favor predatory consumption of other animals, as in lions and tigers. An analogous form of reasoning (i.e., reverse engineering) is applied to *social* traits such as "altruism." Altruism seems, at the outset, an unlikely candidate for a natural selection argument since "good Samaritans" who lay down their lives for others would seem to be selected against in the long run. Unless they reproduced before their acts of self-sacrifice, they would become as rare as hens' teeth. However, under some conditions, altruistic behavior may be adaptive (i.e., genetically selfish). The whistling marmots that let loose their shrill whistles as predators approach the colony may be exposing themselves to individual predation, and, indeed, those closest to the predator may be more liable to be eaten. However, if the majority of the colony survives, natural selection will favor "altruism." This is premised on the condition that members of this closely related group all share the same genetic disposition (to be altruistic), and that in the long run such self-sacrificial behavior permits more marmots to survive than are lost to predators. Biologists acknowledge that such processes probably occur at the level of "kin groups," that is, initially small, genetically homogeneous, interrelated family groups. Evolutionary psychology is premised on the idea that the selection process described for altruistic behaviors in animals provides grounds for inferring how distinctive *social* behaviors could evolve in humans (Tooby and Cosmides 1996).

Some evolutionary psychologists argue that the human brain contains numerous "modules" which evolved to solve certain problems confronted by human ancestors. Kin group altruism would be one of many. Among those identified are: a face recognition module, a spatial relations module, a tool-use module, a child-care module, a grammar acquisition module, etc. (Cosmides and Tooby 1992:61). This implies that the architecture of the brain is constituted by bundles of modules each of which has a distinct neural substratum that is heritable. Such "massive modularity" approaches have attracted sharp criticisms because they seem to imply that gender and racial attributions are fixed modules (i.e., natural), and that the theory plays into the existing power structures of society (Grossi, Kelly, Nash, and Parameswaran 2014). These functional specializations of brain architecture are simply speculation (Chiappe and Gardner 2012; Grossi 2014).

Resistance to the Darwinian approach to human social behavior

Students of human nature identify a cluster of difficulties in accepting the plausibility of natural and sexual selection pressures when it comes to human social behaviors. Human societies are thought to structure social behavior through cultures that consist of historical, as opposed to genetic, memories. It is argued that culture differentiates humanity from non-human species, so that the major determinants of human social behavior arise from ontogenetic experiences, not phylogeny. In addition, human behavior is voluntaristic, that is, based on free will and agency. Evolutionary psychology implies, in the minds of some people, that we are automatons, or puppets whose behaviors are determined by our genetic programming. Genetic determinism contradicts the entire rational choice foundation of the social sciences from Aristotle to Hobbes to contemporary learning theories. And, finally, the combination of the historical accumulation of beliefs and values captured in cultures at the macro level and the processes of socialization and indoctrination at the micro level makes a science based on genetic mechanisms appear irrelevant or redundant. But are these criticisms well founded?

I return to an observation from Carol Gilligan that strikes a chord with the Darwinian approach. As I mentioned earlier, Gilligan formally disavows a biological foundation for gender differences. What I did not mention is that she was equally critical of sociological determinations of gender differences.

> I find the question of whether gender differences are biologically determined or socially constructed deeply disturbing. This way of posing the question implies that people, women and men alike, are either genetically determined or a product of socialization – that there is no voice – and without voice, there is no possibility for resistance, for creativity, or for a change whose wellsprings are psychological.
>
> ([1982] 1993:xix)

She goes on to say that biological reductionism, as well as sociological reductionism, pave the way for totalitarianism, because they both conceive of social action without reference to "voice." Voice is the expression of individual aspirations and responsibilities. It is the core of the "classical tradition," that is, the idea that individuals are autonomous agents responsible for their own actions. For Gilligan, neither genetic conditioning nor cultural conditioning captures the field of action negotiated by individuals as they muster their resources, opportunities, and desires in everyday life. This position is, ironically, shared by the evolutionary psychologist. How could that be so?

Choices and appetites

Evolutionary psychology does not supersede agency. It does not ignore the role(s) of culture and neither does it replace learning theories with behavioral genetic mechanisms that are indifferent to experience. It deepens our understanding of choice behavior by focusing on the things that characterize our appetites. Bentham ([1789] 1970) allows human actors free will, but the expression of that will is tempered by two masters: pain and pleasure. Various systems of control inhibit the individual's acquisition of pleasure and self-interest: the physical, the moral, the religious, the state, etc. The physical system, for example, inflicts costs on persons who pick fights with opponents larger than themselves. Sexual excesses are inhibited by STDs, gluttony by heart disease, etc. The informal moral system attaches costs in the forms of pains of "conscience" and loss of status in the eyes of the reference group. The formal legal system attaches penalties to transgressions in terms of arrests, fines, and confinement. In the terms of classical economics, people "maximize their utilities" by calculating the balance between costs and benefits, or, in Bentham's terms, between pain and pleasure, as experienced within these systems of constraint (Gottfredson and Hirschi 1990). Evolutionary psychology enlarges this picture by suggesting that human desires do not materialize out of thin air. We have evolved highly discerning taste buds to identify salt, sugar, and protein, and have become more successful foragers as a result. In a parallel way, we have evolved social preferences that we retain as "intuitions" or "instincts" that guide our choices in everyday life. So that, in exercising our choices, evolution may have shaped the things we desire and enjoy.

In *The Evolution of Desire*, David Buss explores gender differences in preferences for a mate. These differences appear to have evolved to deal with the differential costs to males and females that arise from mating behavior. The parental investment of males and females is significant but not equal. The females carry the fetus for nine months, suckle the newborn for months, if not years, and assume a large measure of parental responsibility for raising the offspring. Pregnancy has significant and unavoidable opportunity costs for the female but less so for the male, who could choose to father many offspring simultaneously with different mates. There are many potential evolutionary solutions to this cost differential. Buss argues that, in *Homo sapiens*, this has resulted in female preferences for older, taller, higher status, economically successful, generous, and faithful mates. Buss's evidence was based on a number of cross-cultural surveys "in thirty-seven cultures on six continents … Women across all political systems …, all racial groups, all religious groups, and all systems of mating (from intense polygyny to presumptive monogamy) place more value than men on good financial prospects" (Buss 1994:24–5). In fact, women valued financial resources twice as much as men. "These findings provide the first extensive cross-cultural evidence supporting the evolutionary basis for the psychology of human mating" (1994:25). The same pattern emerges in analysis of personal advertisements placed by men and women looking for partners: women seek older, financially secure partners. The preference for older and higher status males is, according to Buss, a marker for economic security and success. Height is a marker for dominance, which is also related to social and economic success. The fact that women want men who are successful, more mature, ambitious, intelligent, dependable, tall, healthy, and faithful will strike many people as “common sense.” After all, are not these favorable attributes in people in general? Buss’s point is that the shaping of desires will result in choices that are “no brainers” *because* they are instinctual. But what is more persuasive is that male priorities are so different. Where women value conditions associated with material security (presumably due to maternal investment), men value youthful women, physical attractiveness (a proxy for health and fertility), a hip to waist ratio of about 0.7, chaste premarital behavior, and postmarital fidelity – all of which are elements that enlarge male fitness. The older males become, the more they desire increasingly younger women, indeed, “trophy” wives who enlarge their status. As a result, males are more interested in casual sex, have different expectations as to at what point an emotional relationship should become physically intimate, have an inflated view of the ideal number of sexual partners, and have a lower threshold for engaging in casual sex.

> Imagine that an attractive person of the opposite sex walks up to you on a college campus and says: “Hi, I’ve been noticing you around town lately, and I find you very attractive. Would you like to go to bed with me?” If you are like 100% of the women in one study, you would give an emphatic no … But if you were a man, the odds are 75% that you would say yes.
>
> (Buss 1994:73)

If women are by nature “coy” and men by nature “randy,” this derives from the costs of casual sex for each gender. Buss also argues that jealousy has an important evolutionary origin. While female baboons undergo external changes when they become fertile, human female ovulation is “cryptic” or non-obvious. Because the male cannot monitor his mate during ovulation, this makes it somewhat more difficult for human males to be certain of the paternity of their mate’s offspring. In evolutionary history, men whose mates copulated with other males would have undermined their own fitness if they had spent years investing in non-progeny. Buss argues that the emotion of jealousy evolved as a (potential) solution to this problem. Both males and females have a proprietary interest in the fidelity of their mates, but how they experience the loss of bond exclusivity is quite different. Buss reports that when male and female subjects are asked to imagine different kinds of infidelities, ranging from spending time with a sexual rival, giving that rival gifts, or actually engaging in sex with the rival, males were far more agitated by their mates having sex with their rivals, while women were more agitated by their mate’s *emotional* attachment to the rival. Women’s jealousy “is triggered by cues to the possible diversion of their mate’s investment to another woman, whereas men’s jealousy is triggered primarily by cues to the possible diversion of their mate’s sexual favors to another man” (1994:128). This was also reflected in differences in physiological distress measured by changes in heartbeat, skin conductance, and other measures of arousal. Men reacted far more to thoughts of *sexual* infidelity than women, and women reacted far more to thoughts of *emotional* infidelity than men. These patterns were investigated cross-culturally and the same differences in fears of jealousy emerged. “These sexual differences in the causes of jealousy appear to characterize the entire human species” (1994:129).

Buss analyzes specific strategies that reflect each step in the mating game. Under “attracting a mate” he describes how individuals display their resources, commitment, and physical prowess, how they try to enhance appearance, and convey sexual signals to express interest. He also outlines strategies for “staying together” and dealing with “sexual conflict and competition,” and the contribution of fitness concerns in “breaking up” (infidelity, infertility, withdrawal of support, withdrawal of sexual access, etc.). In each case, he describes the differences in strategies for males and females, differences that arise primarily from differences in parental investment, and differences in fitness value over the life cycle. It is clear that each specific strategy that he identifies is not hardwired in the sense that Parkinson’s disease is hardwired, expressing itself ineluctably after the age of fifty and proceeding through a set of steps that ruin the nervous system and make premature death unavoidable. Fitness pressures and differences in parental investment shape desires sometimes in vivid ways – intense male insecurities over infidelity – and sometimes in more generalized ways – where, for example, women come to find attractive a range of social traits because of their *indirect* linkage to the material security that evolution has mandated as a priority. Even though they are thought to have a position somewhere in the human genome, these feelings and desires are not impervious to cultural pressures, and neither are they totally removed from influences of learning and reinforcement.

The point that I wish to emphasize here is that evolutionary psychology does not retire “utilitarianism” or “rational choice theory” or learning theories. It attempts to shed light on the contribution of selection pressures to the evolution of our social priorities in the choices we exercise. In other words, it tries to make intelligible what are, for Bentham, merely *generic* “pains” and “pleasures.” And it makes intelligible many social behaviors that seem to be patently irrational. Buss’s explanations for human strategies for mating in terms of different gender priorities sometimes strike us as all too obvious or commonsensical. The analysis of crime within a Darwinian perspective is another matter.

Explaining murder: “trivial altercation,” polygyny, and status

Jack Katz describes in detail the paradoxical behaviors of the “hard-man” robber. These individuals frequently weave “stickup” into a fabric of other criminal activities that include pimping, assault, narcotics, and gambling (1988:165–6). They cut a flashy figure in criminal circles, buying new clothes and giving gifts to friends, partying at length, and cyclically finding themselves broke as a result. They are also far more likely than other career criminals to spend a great deal of time in jail before they “square up” – typically half their adult lives. They often cultivate fearsome reputations because of their employment of what some have dubbed “recreational violence” – a fact that makes them threatening not only to victims but to other perpetrators. Ironically, the average take from a non-bank robber is small – $200–300. Bank robberies, which have an extremely high clearance rate, net about $2,000. So, "hard-men" into the robbery game go on sprees of stickup, followed by sprees of partying, alcohol and drug consumption, gambling, and whoring. Katz asks what the rationality is of this life style. It's not about the money, since no one has anything to show for it at the end of the day. It's not that crime has become a form of work, since the offenders spend half their adult lives in prison. Katz points to the emotional attractions of the lifestyle, but notes that it does not appeal to everyone.

The first attempt to understand the emotional appeals of robbery and violence from a Darwinian psychological framework was made by Daly and Wilson in *Homicide* (1988). They point out that the most prevalent form of male-to-male killing in contemporary Western societies starts from "trivial altercations." This has been noted by criminologists for generations, but no one offered a credible explanation beyond identifying gender and age as behavioral hazards. Trivial altercations are fights that start over "stupid little incidents," arguments, insults, even accidents. Often the combatants are egged on by their associates and friends, and often the violence escalates to the point where someone produces a knife or revolver and conducts a lethal attack, or has the knife or gun taken from him by an adversary who uses it on him. No one planned the killing beforehand. Usually it is a toss-up as to who will win and who will lose. And, typically, the matter that results in a homicide is "a little old fight over nothing at all." Daly and Wilson point out that the fights mean a great deal to those who pursue them. What is at stake is "face," reputation, or credibility in the eyes of one's associates. Daly and Wilson point out that in pre-industrial societies, violence is an important social commodity that is associated with respect and power in village society, and that men who gain prestige through their willingness to kill, often on the smallest pretext, in fact enjoy greater status as well as greater fitness, that is, more wives, and more children.

How do we get from violence to fitness? Daly and Wilson argue that *Homo sapiens* is essentially a polygynous species. In fact, in the anthropological record, over 80% of societies practice polygynous marriage. Many argue that European societies practice serial polygyny. Polygyny is important in understanding the problem of *fitness variance.* In a polygynous species such as the fruit fly, every fertile female will have offspring, but some male flies will breed a great many times, and some not at all. The female's fitness is limited by the number of eggs she carries. The male is limited by the number of mates. Daly and Wilson draw a parallel between the predicament of people and fruit flies:

> A man – like a fruit-fly – could always increase his expected fitness by gaining sexual access to one more fertile female, regardless of whether he presently has no mates or fifty, whereas a woman – like a female fruit-fly – typically would not enhance her expected fitness by gaining sexual access to every fertile male on the planet.
>
> (1988:139)

As a result, female fitness is largely guaranteed, but not male fitness. This has consequences. In many polygynous species, males fight one another to establish dominance. Elk and deer develop large racks of horns, not for defense against predators, but to establish their dominance over the male competitors in their own herds, and, hence, their right to breed. Where a dominant male establishes exclusive reproductive control of the females, his control is subject to challenge from younger males, often in bloody and deadly contests. Polygynous species are characterized by sexual dimorphism, higher rates of male mortality due to intraspecific conflict as well as differences in male–female rates of senescence (longevity), traits that are found in our species.

If *Homo sapiens* is essentially polygynous, this would explain the cross-cultural patterns of male overrepresentation in homicide and other violent crimes. Men, but not women, find the resort to violence attractive to establish status. And it appears to be the intangible aspect of homicidal altercations and the ostentatious quality of the robber lifestyle that matters. As Daly and Wilson note, if the explanation for robbery were penury, most robbers would be poor, old women (1988:178). But most robbery, robbery-homicide, and homicidal altercations are male-dominated activities, if not a male monopoly. Why?

> Men's minimum needs for survival and sustenance are hardly greater than those of women. And the men … are certainly no more likely to be desperately poor than their female counterparts. But in a paternally investing species such as our own, males gain reproductive success by commanding and displaying resources that exceed their own subsistence needs.
>
> (1988:179)

Their account does not end here. Violence is not the only way to acquire status, indeed, it is among the least feasible in modern societies. In their examination of homicide statistics, Daly and Wilson point out that males kill other males at a rate ten times that of females killing other females, that they tend to kill persons of the same age as themselves (i.e., their competitors), that the age of highest risk occurs during the period of most intense family formation (the early twenties), and, finally, that those who engage in such activities are far more likely, compared to the population at large, to be unmarried and unemployed. The appeal of homicide is greatest for those whose fitness is most precarious: young, poor, and unattached males. This analysis also explains Katz's findings about gender, race, and age in his study of robbery. Just as escalating a trivial altercation to the point of homicidal violence only appeals to males lacking other resources to establish reputation, the appeal of robbery only makes sense to young men living in communities lacking access to the legal avenues of wealth and status acquisition. Note in all this that we are not contesting the fact that these individuals are free agents, or that they are not responsible for their actions. Evolutionary theory is explaining the appetites for ostentatious and violent displays and why they have the distinctive gender, age, and social configurations we find in contemporary studies of murder and robbery. This analysis provides a non-obvious explanation of the underlying rationality of behaviors that are otherwise inexplicable.

Learning theories and culture in evolutionary psychology

Many students of social psychology view the evolutionary perspective as redundant because, they say, gender differences are learned. "It's all a matter of socialization," and, as such, it can be easily changed if people simply choose to raise their children differently. This view falsely juxtaposes nature and nurture. When we try to "shape" a behavior in animals through operant or Pavlovian means, learning theorists point out that there already exists an *unconditioned* reflex (Breland and Breland 1966). Gender socialization is the *modification* of existing dispositions, which are not learned, but which appear developmentally. The infant is not a *tabula rasa* at birth, but comes equipped with certain dispositions that are natural and that can be reinforced. The learning of language is the paradigm example. People have a genetic disposition to acquire speech. Cultures may differ in which language is imparted, but the underlying ability to acquire speech is genetic. And it is also developmentally sensitive – if the child fails to receive instruction in the first decade of life, the ability to acquire speech subsequently is severely impaired. There are parallels to gender. Gender differences occur as early as we can measure them. Boys are far more likely to be born prematurely, to show deficits in motor and social behavior, difficult temperament, hyperactivity, emotional disorder, and aggressiveness relative to girls (Pevalin, Wade, and Brannigan 2003). They are also *more* likely to have parents who are hostile and depressed. Socialization is not a one-way street, and it does not work on a blank slate. Evolutionary psychology acknowledges that pain and pleasure can help *shape* behavior, but learning theory is falsely seen as a substitute for a Darwinian approach to the understanding of behavior. Skinner's rats learned to press bars and run mazes to get fed, but feeding was already an innate trait, and the experiment capitalized on its evolved versatility. As Breland and Breland (1966) point out, Skinner's experiments reflected the innate abilities of his experimental subject, the white rat, and reflected the behavioral repertoire found in the rat's ecology. But all the reinforcement in the world could not make the skill sets of rats interchangeable with those of cows, dolphins, pigs, or cats, a lesson the Brelands, who were students of Skinner, learned in the course of careers trying to condition some 8,000 animals representing sixty different species. The point is that learning theory is not an alternative to the evolutionary perspective. The social construction of reality is erected on a biological foundation.

What about culture? Buss explicitly acknowledges that cultures have an enormous influence on the expression of evolved appetites. "Cultural conditions determine which strategies get activated and which lie dormant" (1994:15). Where social welfare programs cover costs of child care, maternity benefits, and material support, as in Sweden, the value placed on premarital chastity declines. “Women’s economic independence from men lowers the cost to them of a free and active sex life before marriage, or as an alternative to marriage. Thus, practically no Swedish women are virgins at marriage” (1994:69). For Buss, the evolutionary legacy of human desires is context sensitive. Whatever the “default setting” in terms of desires, individual actors tailor their actions in the face of external conditions (culture) that modify how they are expressed. Again, the point is that culture does not replace evolved appetites, any more than evolutionary psychology replaces utilitarianism. It determines which appetites enjoy expression, and which are conditioned toward extinction.

Problems with the evolutionary approach to gender differences

One of the virtues of Gilligan’s theory of vivid differences in male and female moral outlooks was that it made it easy to expose it to evidence to determine whether it had empirical validity. The problem outlined in this chapter was that she persisted in her views in the absence of objective evidence and the work became elevated due to its extra-scientific appeals. The situation with evolutionary psychology is more complex due to the nature of the arguments found in this area. There are four major points to raise.

First, the nature of the explanations of the instinctual basis of human social behaviors is explicitly a *post hoc* argument. This derives from the reverse engineering approach that is unavoidable when we move from the deductive nature of evolution in the *general* sense to the more inductive application of fitness models in the case of *specific* social traits or instincts. In the general explanation, we suggest that traits (unspecified) vary to some extent randomly, and that, if these confer reproductive advantages, they are preserved in the organism’s genetic legacy. When we move from the general case to arguing for the adaptive advantages of specific traits, the explanation becomes historical, since we are looking for evidence of changes in the fossil record, peculiarities in biogeographic distributions of fauna and flora, and homologies across related species. Goudge (1961) refers to “historical explanations in evolutionary theory” and the role of narratives in framing the conditions found in the historical record. For example, explaining how amphibians developed limbs that allowed them to evolve into land creatures consists “in proposing an intelligible sequence of occurrences such that the event to be explained ‘falls into place’ as the terminal phase of it … Thus the explanation proposed is an historical one” (1961:72). Desmond offers a similar narrative to explain how hot-blooded dinosaurs evolved into birds. In *Archaeopteryx*, it is argued that the feathers evolved as a means of trapping insects in a species that was initially ground-living, fleet of foot, and large brained. “The bird-dinosaur, complete with endothermy, feathers, wings and a brain able to coordinate intricate manoeuvres, was *completely* ‘preadapted’ to flight” (1975:175).

The same historical problem faces the explanation of social traits. For example: “If, over evolutionary time, generosity in men provided these benefits repeatedly and the cues to a man’s generosity were observable and reliable *then* selection would favor the evolution of a preference for generosity in a mate" (Buss 1994:21, emphasis added). Or, sperm count in male ejaculate increases significantly if a couple spends time apart. "This increase in sperm is precisely what would be expected *if* humans had an ancestral history of some casual sex and marital infidelity" (1994:75, emphasis added). The presumption is that the trait actually enhanced fitness and was heritable. This leads the theorist to speculate on how such an advantage would work. But this is clearly speculation. Stephen Gould (1978) compared such accounts to Kipling's "Just-so" stories, such as how the leopard got its spots or how the tiger got its stripes.[3] The psychologist's reconstruction trades on prehistorical ecological histories that are imagined, since usually the only clue to the fitness pressures are the outcomes and their "design features," that is, the thing we are trying to explain. Sometimes, good use is made of comparative methods, as in the examination of "sperm wars," where it is possible to relate the size of primate testes (in chimps, gorillas, and humans) to differential demands of mating, but often this comparative evidence is lacking (Harcourt, Harvey, Larson, and Short 1981). The method of "reverse engineering" requires us to accept the conclusion first, and search for the evidence later. In organic evolution, we start with the morphological peculiarities of a bone, feather, or hair specimen and conjure the ecological pressures that would be necessary to create them, but in the realm of social behaviors we are not even sure that that the peculiarities we identify (in contrast to organic traits) are heritable.

This leads to my second point. The theory of truth in such explanations is quite different in this area of psychology than elsewhere. The experimental method, in principle, is based on a test of the null hypothesis. Statements of relationships are made in hypotheses, and evidence is marshaled to determine if the relationships are as predicted. The evolutionist proceeds by putting together pieces of a puzzle without knowing what the original pattern looked like and without necessarily having all the pieces. The test of the theory is its coherence or integrity. It makes sense of things that otherwise strike us as unconnected or incoherent. The problem is that it is logically possible for several alternative puzzle solutions to be equally coherent. One is reminded of Sigmund Freud's ([1957] 2001) psychoanalysis of Leonardo da Vinci. Freud saw in Leonardo's painting of the Madonna and Child with St. Anne a disturbing subliminal image: a vulture disguised in the folds of the mother's clothing that threatened the child. Leonardo had recalled from his youth that he was attacked by a bird, specifically a kite, whose tail brushed his lips. In Freud's sources this was mistranslated as a vulture. From these fragments, Freud deduced Leonardo's aversion to his mother, his homosexuality, and other themes of his paintings, including the ambiguity of Mona Lisa's smile. He wove together a coherent account of Leonardo's life that reconciled the tensions and contradictions in Leonardo's life. The problem with the analysis was that the painting Freud analyzed was not actually painted by Leonardo. Typically, master painters would compose the main subjects and leave details such as clothing to their assistants, but, in this case, the master appears to have lost interest in the work before it was completed. So the story Freud pulls together may be coherent, but its empirical premises may be dubious. Coherence may be a necessary condition of truth, but it is not sufficient.

The third, and most worrisome, aspect of an evolutionary account of social behaviors is that cultures can *mimic* traits that may have an instinctual basis. This is a point advanced by Daniel Dennett (1995), a philosopher highly sympathetic to Darwinian thinking.[4] Just as sexual selection, for example, can create an appetite for status through escalation of violence, as discussed earlier, this "trait" can be reinvented by *observers* of trivial altercations, and embedded in social histories that effectively influence people who are unaffected by heritable dispositions for face-saving violence escalation. In the result, young men may adopt aggressive ego contests that result in violence to establish reputation. In effect, an inherited trait may be independently re-engineered in a cultural group as an adaptive strategy that is wholly cultural. Machismo cultures may shape male appetites for violence by rewarding conflict with status, which may, in turn, influence fitness. "With the human species, as Dan Dennett observed, you can never be sure that what you see is instinct, because you might be looking at the result of a reasoned argument, a copied ritual, or a learned lesson" (Ridley 2003:55). The implications of this cannot be overstated. What it suggests is that the entire field of evolutionary psychology consists of the *identification* of hypotheses, often through brilliant "reverse engineer" reasoning. Establishing the evidence for a hypothesis to choose between a model of culturally based fitness as opposed to organic fitness is another matter. I am unaware of any principled attempts to put "paid" to Dennett's position.

The final point concerns the moral ambiguities in the evolutionary psychology perspective. Sometimes, the authors write as though the social behaviors they explain are *vestigial*, in the way that the human appendix is a vestige of a second stomach and functions quite differently from the "main" stomach. For example: "The man who hunts down and kills a woman who has left him has surely relapsed into futile spite, acting out his vestigial agenda of dominance to no useful end" (Daly and Wilson 1988:219). Or consider Buss's reflections on whether he should have published the bad news about men's preoccupation with young, fertile females.

> Suppression of this truth is unlikely to help, just as concealing the fact that people have evolved preferences for succulent, ripe fruit is unlikely to change their preferences … Telling men not to become aroused by signs of youth and health is like telling them not to experience sugar as sweet.
>
> (1994b:71)

Buss also holds that "whereas modern conditions of mating differ from ancestral conditions, the same sexual strategies operate with unbridled force" (1994:14). If they operate with *unbridled* force, why should we have any optimism that "only by understanding our evolved sexual strategies … can we hope to change our current course" (1994:14–15)? Or, "We are the first species in known history of three and a half billion years of life with the capacity to control our destiny" (1994:222). If the sexual strategies are Dawkins-like cultural patterns (memes), this may be so, but if they are instinctual (heritable), we are no more able to deny the "unbridled force" of such appetites than we are able simply to decide that sugar is un-sweet. Evolutionary psychology appears to classify harmful instincts as *vestiges* that can be rooted out by an act of will while the other instincts – our sense of empathy, our ear for music, talent for numbers, linguistic ability – can be preserved at will because, by implication, they represent our good nature. But Mother Nature is indifferent to this kind of romantic thinking: one species' vestige is another species' true nature. It is all the same to the DNA.

Final words

In previous chapters, we have seen that social psychology is frequently influenced by extra-scientific agendas lurking behind the experimental design. In our discussion of gender – from Gilligan to Buss – the experiment *per se* has receded from view to be replaced by anecdotal evidence, but the political agendas have not. In neither case can the science that each advances contribute positively to scientific progress. Where Gilligan has claimed that the political has become the psychological, the evidence reviewed here is quite the opposite: the psychological has become political. Mednick captures this in her analysis of the gender bandwagon in contemporary psychology, as does Sommers in her report on the academic war against boys. Gilligan's case fails because the evidence for differences in moral outlook is all gainsaid. Ironically, evolutionary psychology acknowledges these kinds of gender differences. It acknowledges further that gender conflicts, if not normal, are virtually unavoidable if we recognize the competing interests of men and women, but for different reasons. Men are disconnected because they are competing for fitness. Yet, both proponents write in romantic terms about deciding to "make it right." Both Gilligan and Buss are optimists. If, on self-reflection, one realizes that society and history have over-structured one's experience, one can choose to act differently. One can recover one's "voice." One can say "no" to polygyny. One can say "no" to patriarchy. These may be interesting moral points, but they are not very compelling scientific points.

Notes

1 In 2019, the National Inquiry into Missing and Murdered Indigenous Women reported on the results of its thorough inquiry into missing and murdered aboriginal women in Canada. Earlier police reports by the RCMP had suggested that, in the past several decades, several thousand aboriginal women had disappeared from their reservations and homes, and over 1,000 had been murdered, according to police occurrence reports. Many of their lives were touched by narcotics abuse and prostitution. When investigators from the commission began to explore the record, they rejected the gender bivalence which presupposed that those at risk were easily distinguished by sex. In the 100-page executive summary they refer over 300 times to the group of

interest as woman, girls and *2SLLBGTQQIA*. The latter refers to individuals who are known as Two-Spirit, lesbian, gay, bisexual, transgender, queer, questioning, intersex or asexual (Canada 2019).

[2] The effect size is based on Cohen's *d*-statistic which calculates the standard deviations between the means of the test and control groups (males versus females in this case). A score of zero indicates no difference at all. A score of ± 0.2 is considered small, ± 0.5 is moderate, and ± 0.8 is large. A simple *t*-test of mean differences is misleading with samples of several thousand cases because they may provide statistical differences which are substantially meaningless. Cohen's *d* provides a more meaningful comparison.

[3] This was clearly unfair, inasmuch as the explanation of any specific organic trait would require the sort of narrative explanation of the kind Goudge describes. Gould's position is paradoxical since, as a paleontologist, he clearly subscribed to evolutionary theory, although the variant he proposed as one of "interrupted equilibria" – based on his analysis of the Burgess Shales, and the stunningly varied forms of life in the Precambrian period (Gould 1989).

[4] The idea of "memes" was first introduced by Richard Dawkins in the *Selfish Gene* ([1976] 1990) and elaborated by Daniel Dennett in *Darwin's Dangerous Idea* (1995). Memes are basically memories that reproduce something that confers an advantage. They are the cultural equivalent of genes, but permit the replication of a structure dramatically faster than organic reproduction and permit the acceleration of cultural evolution with increasingly diminishing input from genetic information.

9 The failures of experimental social psychology in the classical period

Why has classical social psychology failed?

In this work we have reviewed some of the classic studies of social influence in the early days of the field, from Sherif to Asch to Milgram and Zimbardo. We have examined the application of social psychology to industrial production (Hawthorne), school performance (Pygmalion), bystander effects, and mass media effects (aggression). And we have most recently reviewed some of psychology's recent contributions to the study of gender. These do *not* represent an exhaustive sampling of the developments of the field, but they cover materials that would have a very high recognition factor among serious students of the field, and they touch on many of its classic contributions, and include studies that are viewed as the foundations on which the field has developed. My point in this book is to suggest that the scientific achievements in the field, as represented in these studies, are rather modest, frequently misleading, and sometimes downright wrong. Social psychology has failed as a science, not because it has made mistakes, but because it appears incapable of recognizing them as such. What is alarming is that the devotees of social psychology either do not seem to attach much gravity to this situation or do not know what to do about it. Carey, in questioning the acceptance of the Hawthorne effect in the absence of any credible evidence supporting it, suggested that the answer lay in the occupational and professional attractions of the idea. He wrote:

> How is it that nearly all authors of textbooks who have drawn material from the Hawthorne studies have failed to recognize the vast discrepancy between evidence and conclusions in those studies, have frequently misdescribed the actual observations and occurrences in a way that brings the evidence into line with the conclusions, and have done this even when such authors based their whole outlook and orientation on the conclusions reached by the Hawthorne investigators?
>
> (1968:416)

In his discussion of the debate over-operationism in the history of the cognitive dissonance, Rosenwald wrote:

> The main objective of this paper is to explore the incentives of a scientific strategy which delivers a scanty theoretical yield and which lacks firm philosophical supports. The premiss underlying the present discussion is that, whereas operationism can be criticized (or defended) on logical grounds, its persistence in the face of methodological critique and of its disappointing theoretical yield cannot. We must seek answers outside the philosophy of science, for instance, in the sociology of professions and in the explicit and implicit missions which learned disciplines set for themselves.
>
> (1986:303-4)

Mednick argued that *bandwagons* have characterized a great deal of the research on gender and, I might add, media effects. I have argued that in all these cases the scientific agenda has been drawn below the threshold of critical radar by a powerful "tractor beam," that is, by unacknowledged moral and political agendas. These have frequently dictated the scientific agenda and sabotaged the prospects of scientific progress. Ironically, they have also kept the field alive. The idea of the failure of the field of social psychology has to be taken with a grain of salt. The failure that I refer to borrows from previous reports of psychologists discussed in earlier chapters who recognize that the scientific project has gone off the rails. Recall Zajonc's remarks, looking back over the past four decades, that social psychology has not developed any consensus about what is central to the field, nor accumulated any evidence to establish important, non-obvious lawlike statements about human behavior. Recall Buss's characterization of the field as in "theoretical disarray," or G. A. Miller's characterization of the field as "an intellectual zoo" without any standard method or technique. Cooper concluded that psychology's failure to progress was to be attributed "not to immaturity but to retardation." Ferguson characterized the history of psychology as decades of "useless research." These rather grim assessments of the discipline come from people who have enjoyed successful careers in the field. In spite of the concerns I have outlined, the field flourishes as a popular undergraduate subject, and its publications, journals, and conferences proceed as though none of what I argue is valid or relevant. This motivates me to clarify what I mean when I argue that social psychology has failed. My concerns lie in three areas: methods, theory, and ethics. We shall examine each in turn.

Methods

In the area of methods, there are three issues. The first has to do with the logic of experimentation. Psychology's purchase on the scientific community has been made on the basis of its scientific methodology, particularly its emphasis on experimentation. As every student of methodology knows, the gold standard in terms of empirical inquiry is the true experiment: random assignment of subjects to treatment groups, identification of correlations between treatments and outcomes, temporal precedence of the causes over effects, and an ability to rule out spurious relationships. The reality was rather different. At the heart of the discipline covering the period, from Sherif to Asch to Milgram to Zimbardo, the "experimental" projects carried on in the laboratories were primarily "demonstrations." They were undertaken without the intent of explicitly testing specific theories. Sherif *simulated* how norms evolved. Asch *mimicked* resistance to propaganda. And Milgram *dramatized* obedience to authority. The demonstrations were inductive or exploratory. There were no tests of significance. They attempted to paint a picture of realities that were already, to some extent, well understood, at least in a common sense way. The studies of media effects of violent television, video games and pornography through the use of Milgram-type shock designs were not studies of the actual effects of media on sexual and physical aggression, but the imagined *parallels* or *analogs* to what such effects on aggression might look like *if* they were real. So, rather than exposing ideas to the potential of falsification that derives from explicit theory testing, classical social psychology replaced the scientist with a dramatist. It made for memorable stories, but little in the way of theory development.

The second major issue follows directly. Because demonstration was so central to the academic culture of experimental social psychology, there was little evidence of disconfirmation. Indeed, without any explicit theory, the idea of disconfirmation was almost foreign to the research culture. This was also noted by Pepitone in his reflections noted earlier, when he recalled that the vast majority of published research articles' findings confirmed their hypotheses, and that progress through falsification was rare. As we observed, Festinger explicitly questioned the logic of falsification, and, in his own work on cognitive dissonance, the field "progressed" despite the profoundly ambiguous empirical results that it generated. Rather than conclude that the underlying model was problematic, experimenters bent themselves into pretzels to preserve it (Cooper and Fazio 1984).

The final point is that the field appears to have handicapped itself by a virtually exclusive devotion of its research methods to experimentation. The crisis literature that emerged in the 1970s returned to this point repeatedly. The field became wedded to the belief that complex social events could be examined causally in the laboratory, even though such experiments of necessity were of short duration, low impact, and typically drew from a restricted sector of the population – the undergraduate psychology majors. This limitation was noted, but no serious movement to broaden fundamentally the methodological scope of social psychologists ever proved successful.

Theory

Given the preoccupation with methods, it is hardly surprising that social psychology has failed to develop substantial headway in the development of a theory of action. The situation is exacerbated further by the legacy of common-sense psychology, that is, the idea that, at the meso level, most actors already have a relatively sophisticated understanding of interpersonal interaction. Common-sense knowledge and scientific knowledge claim the same territory. By contrast, evolutionary psychology theory has made great strides in our understanding of the foundation of appetites that otherwise strike us as perplexing (i.e., "trivial altercations"), but its insight is not based on methodological premises as much as a larger Darwinian explanation of adaptation – a much-needed, strong, theoretical basis. However, this promising development has yet to attract much attention in the core of social psychology because it does not lend itself readily to experimental testing, any more than Darwinism does in zoology or paleontology. Also, psychology has yet to sort out all the mechanisms of sense-making that are attributed to the evolved brain – cheater detection, nepotism toward blood relatives, altruism, moral compulsions, retributive justice, etc. – and to determine their relative importance *vis-à-vis* the long-standing mechanisms of adjustment – classical and operant conditioning.

In social psychology, this avenue of theory development has been overshadowed by the moral agendas that have grounded the demonstrations of the classical tradition. The most important theoretical development in the field, cognitive dissonance, did not prove successful, but, rather than scrubbing the agenda and moving on, a generation has attempted to redefine and finesse it, as though incapable of acknowledging that not every new idea is scientifically sound. The demonstrations of Asch, Milgram, and Zimbardo, as well as the field studies in industry (Hawthorne) and schools (Pygmalion) were landmark *moral* achievements, but they did not advance psychological theory. George Miller (1992:40) warned that "pandering to public interest" would destroy the scientific integrity of psychology. While this may be an overstatement, in my view, the classical tradition, because of its "relevance" and extra-scientific agendas, has suffered in theoretical development.

Ethics

The final area of concern is ethics. As recounted in previous chapters, many of the classical studies were quite provocative and, in some cases, traumatic for the human subjects. Let us examine some cases. Zillmann and Bryant (1982, 1984) recruited 160 students for a study of the effects of pornography on attitudes toward women. One group viewed a menu of sexually explicit films for one night per week for a period of six weeks. Subjects were tested in three further weeks for evidence of calloused attitudes toward rape, rape myth acceptance, tolerance of censorship, and sympathy for the "female liberation movement." Changes in attitudes were found in both male and female subjects. The authors suggested that participation in the experiment had produced "non-transitory" shifts in attitudes toward rape myth acceptance, increasing indifference to victims of rape, etc. If we accept that changes for the worse were "non-transitory," were subjects' attitudes "injured" as a result of this study? I do not believe that the changes were permanent, contrary to what the authors suggest, but, as experimenters, are we not simply gambling that the risks associated with experiments are always tolerable? To their credit, Zillmann and Bryant suggested that there should be an embargo on future pornography effect studies since the harm to subjects had been established in their eyes. Ironically, this would have the effect of curtailing the replications that were so unsuccessful in corroborating their findings.

Consider the drug emotion studied undertaken by Stan Schachter with his students in the early 1960s. In order to control levels of arousal to a film that was calculated to produce emotion (i.e., laughter – the film was Jack Carson's *The Good Humor Man*), subjects were injected with a powerful sedative – chlorpromazine. Ladd Wheeler explains that the graduate students had to experiment on themselves to determine a suitable dose to administer (blind) to undergraduate subjects.

> We pretested the chlorpromazine doses at 50 mg on ourselves and other graduate students, and Stan had us make notes of our feelings. Chuck Hawkins wrote that he had decided he was definitely going to die, after he checked his pulse at 32 and falling. Bibb Latané came out of the testing room and promptly fell on his head, knocking over the coffee pot. Stan consulted all sorts of experts and finally decided to halve the dosage, in the face of overwhelming ignorance on the part of the experts. Mental hospital patients are given extreme dosages, but no one knew what it might do to an undergraduate. Even then, we had a cot available for the chlorpromazine subjects, and it was used with some frequency after the experimental session.
>
> (1987:48)

Wheeler goes on to say that they had the welfare of the subjects as a first priority and that a physician was always on hand. But how could anyone determine a "reasonable" dosage when *expert* knowledge could provide little guidance, and when the primary clinical use of the drug was to treat schizophrenics? And what sort of human experiment requires a *frequently used* cot to help subjects recover? Again, I have no reason to believe that anyone suffered serious injury in these experiments, but were the psychologists not simply taking risks with the adjustments of their graduate and undergraduate students? Zimbardo (1999) argued that the Institutional Review Boards (IRBs) had overreacted to the treatment of human subjects in psychological experiments. When we see the nature of the interventions undertaken in this period, can we honestly claim this was overreaction? Milgram's subjects sweated, trembled, stuttered, bit their lips, groaned, dug their fingernails into their flesh, and experienced uncontrollable nervous laughing fits and full-blown, uncontrollable seizures. This was the most important "experiment" in the classical tradition.

Some sense of the lack of concern over ethical treatment of subjects is suggested indirectly in Leon Festinger's reflection on the issue of deception and ethical treatment of human subjects. He refers to the Tuskegee syphilis study, in which 399 American black men, poor sharecroppers, in Macon County, Alabama, were misled by the U.S. Public Health Service about their illness. They were not treated for their infections, but were simply monitored over a period of four decades (1932–1972) and developed grotesque symptoms that could have been eliminated by antibiotics and sulfa drugs. As Festinger notes:

> one group was treated; the other was also followed in time but not treated. Was this an ethical violation? Is it "harming" someone, who would have had no medical attention anyway? .. *I'm not so sure about this.* These persons were not harmed by the research in the sense that they were no worse off than if the research had not been done at all.
>
> (1980:249, emphasis added)

This opinion seems ethically challenged on several counts, not the least of which was the secrecy that made the subject participation uninformed, and the medical experimenters' abandonment of the Hippocratic oath – *do no harm* – even by omission. One has to wonder not whether, but to what extent, the current restrictions on the protection of human subjects owe their origin to the insensitivity of earlier "masters" to issues of ethical treatment of human subjects.

Social psychologists quite properly note the overreach that has accompanied the new ethical environment that holds psychological researchers to the same standards that govern medical research. The only justification for exposure of human subjects to an experimental drug therapy is that it is undertaken to improve their health. As a consequence, obtaining approval in the new ethical environment for the sorts of deceptions carried out in the classical period is quite difficult. In his replication of the Milgram study, Burger (2009) limited the maximum shock level to 150 shocks, since this was a level that corresponded well to the pattern of people who administered all 450 volts. Burger was able to get comparable results without the trauma that marked the initial experiment. Despite what I think were dubious practices in the past, I am not certain that the new regime of subject protection is wholly desirable. For example, because of the IRBs, a student who writes a story about coaches and sports violence for the student newspaper is less encumbered than the individual who investigates the same issues for a research paper. A professional researcher is more encumbered by ethical strictures than a professional journalist who may be working on the same story. The professional researcher faces significant limits of action from committees struck in the first instance to review *medical* experiments on human subjects.

Consider the predicament of Elizabeth Loftus and Mel Guyer. A clinical case reported in *Child Maltreatment* by David Corwin in 1997 was cited in a court decision to establish the reliability of "recovered memories" of childhood sexual abuse. Loftus (1993, Loftus and Ketcham 1994) had previously established the difficulty of crediting the veracity of repressed memories, and wanted to investigate the case of "Jane Doe" to determine if her recovered memories were as reliable as had been reported. Loftus and Guyer succeeded in identifying the case in court records and interviewed Jane Doe's mother, who, in their view, had been *falsely* accused of sexual assault, and subsequently lost custody of her child. Guyer contacted the IRB at the University of Michigan to establish that their work constituted a comment on a forensic issue that was beyond the framework of the IRB committee. Initially, in 1998, the committee agreed but subsequently it informed Guyer that his research was “disapproved” and that he faced a reprimand for conducting it. After a year of appeals and reviews, the IRB determined it lacked jurisdiction because it was not “human subjects research.” Just as they thought they had the “all clear” signal, Jane Doe sent an e-mail to the University of Washington complaining that Loftus’s inquiry into her case was invading her privacy. Carol Tavris picks up the story:

> On September 30,1999, having given Loftus 15 minutes’ advance notice by phone, John Slattery of the University of Washington’s “Office of Scientific Integrity” arrived in Loftus’ office, along with the Chair of the psychology department, and seized her files. She asked Slattery what the charges against her were. It took him five weeks to respond, and when he did he had transformed Jane Doe’s “privacy” complaint into an investigation of “possible violations of human subject research.”
>
> (2002)

Loftus was eventually cleared but not before her institution had reprimanded her for employing methods inconsistent with ethical principles of professional psychologists, that is, journalism. She was also forbidden to make further contact with the principals in the Jane Doe case. Despite the U.S. Supreme Court’s vigilant protection of first amendment rights, the IRB bureaucracy undertook a campaign of secret accusation, professional harassment, obstruction, and intimidation of leading scholars to suppress their free speech in the name of protecting human subjects. A journalist would have faced none of these obstructions, and would have been better equipped to publish the truth. In my view, this situation is absurd, and it materially threatens the future growth of social psychology. If unchallenged, budding young social scientists would be better counseled to seek a career in journalism.

Outcomes

I have reviewed three areas where social psychology has disappointed – methods, theory, and ethics. The field continues to enjoy great popularity in university courses, but the rise of the IRBs threatens to curtail future research activities, both experimental and, if the Loftus–Guyer episode is any guide, non-experimental. The result of this situation is a lack of progress in the field, recurrent crises that are never convincingly resolved, no basic change in methodological outlook, no theoretical resolution, and volumes of writing that, in the words of Rod Cooper, lead to no new knowledge.

> Every year psychologists turn out thousands of books and articles. I find it difficult, however, to see much in the way of fruits from these labors. In spite of the tremendous publishing record of the psychologists I can see no psychological contributions that I can call marvelous. Think about it. Suppose tomorrow that all these scholarly efforts of the psychologists should

> disappear from the collective knowledge of mankind. … Would it really make a difference? I suggest not.
>
> (1982:264)

When a crisis emerges, it results in professional fragmentation as new subdisciplines of "humanistic" psychologists, feminists, Darwinians, and poststructuralists depart from the experimental core. The field loses its grasp on the larger picture of how all the elements fit together and there is no sense of unity or integrity that forges a coherent discipline. This brings us back to Zajonc's reflections in an earlier chapter: there is no consensus about the key contributions of the field, textbook chapters can be reshuffled randomly without costs because there is no inherent order in the subject matter. Indeed, textbook authors appear unconstrained in representing profoundly important deficiencies in the works they report. Any fair-minded observer would conclude that this marks the fall of classical social psychology as a distinct academic enterprise, and the obituary has been written by its practitioners. As we turn to the contemporary situation, we find that experimental social psychology has entered a new period of crisis surrounding the issues of scientific fraud, questionable research practices, and a wholesale wave of failures to replicate findings. We turn to that problem in contemporary social psychology now.

10 The replication crisis

Social psychology in the age of retraction

Scientific fraud and questionable research practices

Contemporary social psychology has been seized over the past years by a loss of credibility and self-confidence associated with scientific fraud and unsuccessful attempts to replicate the modern corpus of knowledge. We start with fraud and the slippery slope associated with it – the normalization of questionable research practices in contemporary social psychology.

In August 2011, three graduate students approached the chair of the Psychology Department at the University of Tilburg, Professor Marcel Zeelenberg, at a conference in England regarding their suspicions about the datasets created by Professor Dietrick Stapel, then Dean of the School of Social and Behavioral Sciences. They suspected the data in them had been fabricated. Stapel had been recruited to Tilburg in 2006 after a successful career as a professor at the University of Groningen from 2000 to 2006. He had earlier successfully defended a dissertation at the University of Amsterdam in 1997. He was one of the most accomplished researchers in the Netherlands with scores of articles in the leading international journals in social psychology to his credit, as well as awards for career achievements from professional associations. He had a reputation as a hard-working, intelligent scientist and sociable colleague, and was a prolific doctoral supervisor. Zeelenberg brought the allegations of fraud to the attention of the university rector. Both confronted Stapel with the allegations. He initially denied them. Shortly thereafter, he acknowledged that many of his papers contained fraudulent data but he was unsure when he had started making them up. He was suspended in September, 2011 and dismissed shortly thereafter. Committees were struck at each of the three universities with which he had been associated from 1993 to 2011 to determine the extent of the fraud: Amsterdam (the Drenth Committee), Groningen (the Noort Committee), and Tilburg (the Levelt Committee). They published a preliminary report in October, 2011 which created an international scandal. One of the leading lights in international social psychology was a fraud. The final report, *Flawed Science: The Fraudulent Research Practices of Social Psychologist Diederik Stapel,* appeared a year later (Levelt, Noort, and Drenth 2012). Stapel (2012) published his own confession in Dutch in 2012 under the title, *Ontsporing* ("Derailment") (see Brown 2014).

With the assistance of statisticians, the committees examined all the research Stapel had published when associated with each of the three universities as well as the doctoral dissertations he supervised or co-supervised. The committees identified a total of fifty-five journal publications in which "fraud was *established*" including thirty-four while he was at Tilburg (as well as three dissertations), and twenty-one publications while at Groningen (as well as seven dissertations and four book chapters). The Noort committee at Groningen found, not that "fraud was *established*", but that there was "*evidence* of fraud" in another three refereed publications and four book chapters. The Drenth Committee at Amsterdam found "*evidence* of fraud" in seven refereed publications, as well as a *strong indication* of fraud in two chapters of Stapel's dissertation.

> The Committees' findings show that fabrication of data in one form or another started before the Tilburg period. The first publication in Groningen in which fraud *has been proven* is from 2004, and the first publication where *evidence* of fraud was found dates back to 2001. During the Amsterdam period the first publication where *evidence* of fraud was found was from 1996.
>
> (Levelt et al. 2012:25–31, emphasis added)

In the Amsterdam period, "evidence of fraud" referred to inappropriate data collection and analysis, omitting conditions and variables, editing observations and repeating experiments until effects met expectations (2012:32). The later cases at Groningen and Tilburg appear to reflect the simple manufacture of data from ghost subjects who were never actually tested.

Stapel's methods were fairly sophisticated. He discussed potential research topics, typically in the area of priming that looked promising in terms of the existing literature and worked out a detailed methodology including questionnaires, measurements, and experimental protocols with his potential collaborators – graduate students and research colleagues. He would then undertake to run the experiment and implement the data collection himself. Finally, he would turn over the data files to the colleagues for them to begin the interpretation of results. The investigating committees concluded that the graduate students had no culpability in the manufacture of evidence. The record showed that the graduate students never analysed the raw data collected from Stapel's questionnaires or engaged in any debriefing of subjects, although they may have collected other data as part of their research. Stapel's colleagues sometimes commented that his results "were too good to be true" but no one challenged him, including his post-docs and junior colleagues who should have exercised more due diligence. The universities did not penalize any of the co-authors and neither did they invalidate the degrees. The careers of graduate students who were co-authors were already injured when the investigating committees contacted all the journals affected and asked for the retraction of fifty-five of Stapel's papers. A retraction is a formal notice attached to the electronic version of the article which notifies potential readers to ignore it. The attempts to replicate so many high profile “achievements” would be a waste of time and money. And the blatant manufacture of data raised questions about the criminal liability of a scientist who misused public research funds to commit scientific fraud.

One of the most poignant questions raised by the review committees was how it was possible for such dubious scientific practices to escape the notice of all the academic reviewers in the high-profile journals, the funding agencies, and the scientific conferences. Many statistical anomalies were identified readily by statisticians who assisted in the review of Stapel’s papers. The committees were forced to conclude that “there is a general culture of careless, selective and uncritical handling of research and data. The observed flaws were not minor ‘normal’ imperfections in statistical processing, or experimental design and execution, but violations of fundamental rules of proper scientific research” (Levelt et al. 2012:47). The culture contributed to the absence of skepticism about Stapel’s extraordinary findings. For example, in one study published in *Science*, Stapel said he examined the relationship between racism and environmental untidiness. During a strike of sanitation workers in the Utrecht train station, he asked white subjects to fill out a questionnaire in one of the rows behind a person seated in the front row. When the person, a confederate, was black, and when the station was messy due to the unsanitary conditions, people chose to sit further away from the black person. Ergo, racism is primed by an untidy environment (Bhattacharjee 2013). This study was a complete fabrication. A claim so lacking in common sense was accepted at face value.

Stapel has been described as one of the biggest con men of the last decade, but not the only one. Michael LaCour, a graduate student in political science from UCLA and Columbia University professor, Donald Green, landed another work of fraud in *Science*: “When Contact Changes Minds” (2014). They claimed that it was possible to change the opinions of straight people towards an acceptance of gay marriage by having gay canvassers talk to people about their political views at their front doors. The gay canvassers were more effective than straight canvassers in changing attitudes, and the effect was still evident months afterwards. The results were reported widely in the media. David Brookman and Joshua Kalla’s attempt to replicate the results was unsuccessful. When they turned to the company used by LaCour to conduct their survey, they learned that it had never been retained by LaCour in the first place. In their attempts to replicate the study, they noticed a series of statistic irregularities that led them to believe the data were simply made up (Brookman and Kalla 2015). Professor Green was ignorant of the scheme, and Princeton rescinded a job offer to LaCour (Aschwanden and Koerth 2016). The article was retracted in 2015.

Uri Simonsohn (2013) re-analyzed the published data for two psychologists whose statistical analyses of means and standard deviations looked anomalous. When he obtained the raw data, he was able to prove that the chances of the reported effects, after ruling out benign explanations, were astronomically improbable. One researcher, Lawrence Sanna, was studying “embodied morality” – the idea that people behave more altruistically if they are physically elevated, as in riding up on an escalator. Sanna's data showed strong effects for different conditions in the means but no variations in the standard deviations. Analysis of the raw data yielded more irregularities. Sanna resigned his position at the University of Michigan and asked that three of his papers be retracted from the *Journal of Experimental Social Psychology.* Dirk Smeesters of Erasmus University School of Management studied how different colors could lead subjects to either *assimilate* (i.e., identify themselves) with exemplars and stereotypes (blue) or *contrast* (i.e., differentiate themselves) from exemplars and stereotypes (red). The means were all as predicted, but exceedingly so. More problems were identified in the analysis of the raw data, including the unacknowledged omission of data. And all the data (both digital and hard copy) unaccountably went missing. Smeesters resigned in 2012 and three of his papers were retracted from publication. Simonsohn's point was that if the raw data had been posted, these errors would have been evident to other researchers working in the same area. Suspicions of fraud continue to haunt experimental studies of priming (van Kolfschooten 2015).

A rather different case concerns the "critical positivity ratio" published by Fredrickson and Losada (2005). In an effort to demonstrate that the new field of positive psychology had a firmer empirical base than the older humanistic psychology, Fredrickson and Losada claimed that they had discovered a critical ratio based on differential equations and non-linear dynamics which applied universally across persons and institutions, and demarcated the ratio needed objectively to achieve happiness. The tipping point between the ratio of positive to negative experiences that unleashes positive emotions was identified as 2.9013 positive to negative experiences. This applied to individuals and couples, businesses, education institutions, health care systems, marriages, etc., independent of other factors (Friedman and Brown 2018:241). An analysis of the mathematics alleged to underlie the ratio by Brown, Sokal and Friedman (2013) led to the following conclusion:

> We find no theoretical or empirical justification for the use of differential equations drawn from fluid dynamics … to describe changes in human emotions over time; furthermore, we demonstrate that the purported application of these equations contains numerous fundamental conceptual and mathematical errors.

Fredrickson and Losada's claim to have demonstrated the existence of a critical ratio "is entirely unfounded." This critique also led to a retraction of the original paper, but the proponents of the magic ratio (often rounded to 3.0) continued to tout its relevance. This is not a case of scientific fraud, but wishful thinking among proponents of the new directions in positive psychology. It is more like a religious belief than a scientific discovery.

We tend to think that there is a sharp line between outright fraud and the "massaging" of data. Stapel and Smeestra did both, but that part of their publications in which they engaged in grey-line data manipulation appears to be common. The Netherland committees of inquiry into Stapel were told that "this is what I learned in practice; everyone in my research environment does the same, as does everyone we talk to at international conferences" (Levelt et al. 2012:48). Smeestra reported similarly about the generality of data massaging in his area (Vonks 2012; see also BPA 2011). John, Lowenstein, and Prelec (2012) examined questionable research practices in a more general way. They conducted an electronic survey sent to nearly 6000 researchers, including over 2000 psychologists, to survey the prevalence of the use of self-reported questionable research practices (QRPs). What did they learn from the psychologists? One in ten respondents admitted to having falsified data, 67% reported they selectively reported results that "worked," 74% failed to report all their actual dependent variables, 71% reported that they continued to collect data until they achieved a significant result, 54% reported unexpected findings as having been hypothesized beforehand, and 58% excluded data to enhance the significance of their findings. The highest levels of self-admissions of QRPs were found among social psychologists (40%), followed by cognitive scientists (37%) and neuroscientists (35%) (John et al. 2012:530). The fact that 10% of the respondents falsified data is extremely worrisome. In experimental studies when the level of statistical significance is set at $p < 0.05$, everyone accepts that up to 5% of publications will be false positives, that is, accepted even though the relationship is a random effect and probably untrue. But the results from this survey suggest that another unspecified percentage will be irrelevant, since the findings have been manufactured in some fashion in another (as reported by 10% of John et al.'s respondents). The fact that "only" 67% of researchers reported studies that "worked" is more understandable. Researchers do not have incentives to report studies that proved fruitless, and there are few significant venues that are designed to report such work. However, the fact that 71% basically designed their studies to terminate only when they reached a publishable conclusion by adding more cases is basically an admission of gambling with the odds. And the fact that the majority excluded data (58%) or suppressed dependent variables that were actually used (74%) is contrary to basic scientific methodology and scientific integrity. It will play havoc in attempts to replicate because it represents a gulf between what appears in the publication for distribution to the scientific community and what was actually done in the laboratory.

The objective of publication is to make a point, especially a provocative one. For example, Daryl Bem's work on the paranormal led to identification of "precognition" abilities (i.e., evidence of subjects' foreknowledge of the future) in nine different experiments involving over 1,000 subjects (2011). His own attempts to replicate subsequently failed, but he made an interesting admission about his use of experiments:

> If you looked at all my past experiments, they were always rhetorical devices. I gathered data to show how my point would be made. I used data as a point of persuasion, and I never really worried about, *'will this replicate or will this not?'*.
>
> (quoted in Vyse 2017)

That view does not appear to be out of line among psychologists with the high level of QRPs identified in the John et al. survey. And it reflects the approach in classical social psychology where experiments were designed as demonstrations.

Will this replicate? The new crisis in experimental social psychology

There has been a high level of concern both among scientists and in the public domain about the failures of replications associated with important experimental work in psychology. For example, the matter has been discussed in *The Economist* ("Trouble at the lab," 2013; "Try again", 2015), *The Atlantic Magazine* ("Psychology's replication crisis is running out of excuses" Yong: 2018) and *Slate* ("Why psychologists' food fight matters," Meyer and Chabris: 2014). It has also been the subject of special editions in the journals *Perspectives on Psychological Science* (Pashler and Wagenmakers 2012) and *Social Psychology* (Nosek and Lakens 2014). It was the subject of a news release by the American Psychological Association ("A reproducibility crisis?" 2015a). The Center for Open Science has coordinated some large-scale replications of important experiments. Brian Nosek, with 269 other researchers, repeated the experiments described in 100 original well-known papers that were published in three leading psychology journals in 2008. Only thirty-nine out of the 100 replications came to the same conclusions. Even here, the effect sizes of the causal variables were just half what was originally reported. Of the sixty-one studies that were not replicated, some twenty-four produced findings that were "moderately similar" to the original studies but fell short of the minimal statistical significance. The researchers concluded that "39% of effects were subjectively rated to have replicated the original result; and if no bias in original results is assumed, combining original and replication results left 68% with statistically significant effects" (Open Science Collaboration 2012). That puts a question mark over 32% of 100 papers in leading journals. It does not mean that 32% of publications are untrue, false, or fraudulent. It means that almost a third of the publications are not robust or reliable, and there are probably numerous reasons for this.

A smaller scale replication was undertaken to investigate "variability in replicability." This was a unique approach to replication that used a kind of "crowd-sourcing" by recruiting potential replicating scientists through an appeal posted online. Also, it did not involve the replication of specific studies, but clusters of studies that were examining the same effects employing relatively simple research designs. Specifically, the study tested thirteen classic and contemporary effects across thirty-six independent samples that involved 6,344 subjects. Ten effects replicated consistently, one showed weak support for replication and two effects did not replicate. "Most of the variation in effects was due to the effect under investigation and almost none to the particular sample used" (Klein et al. 2014:151). In other words, where there was a strong relationship, it was relatively easy to replicate it. Where the original effect sizes were small, and the levels of statistical significance were most permissible (i.e., $p < 0.05$), that was another matter.

One of the most unique large-scale replications in recent years focused exclusively on experiments that had been published in *Science* and *Nature* – the most prestigious scientific journals in the US and the UK, respectively, which have a reputation for publishing work that is exciting, innovative, and important. A team of twenty-one researchers associated with Brian Nosek and the Social Sciences Reproducibility Project (SSRP 2016) attempted to replicate well-known studies published between 2010 and 2015. The replications were undertaken after consultation with the original authors, and the experimental protocols were pre-registered. The replications typically involved sample sizes five times those of the original studies. This was to give them stronger statistical power (i.e., the ability to find real effects, however small). The results showed that the replicators were successful in discovering significant relationships in thirteen (62%) of the original twenty-one studies. The effect size of the replications was only about 50% of the effects originally published (Camerer, C.F. and twenty-three others, 2018). One of the strongest covariates of failures to replicate was the permissive level of statistical significance employed ($p < 0.05$) and small effect sizes in the original studies. These studies were not a cross-section of social psychology experiments. Their novelty may have led editors in such high-level publications to overlook their reliability. On the other hand, they represent the most prestigious fruits of psychological research and appeared in apex journals.

Another fascinating aspect of the study is that the authors recruited 206 volunteers (psychologists, economists, and graduate students) to place bets on which research papers would replicate successfully before the results were published.

> Each started with $100 and could earn more by correctly betting on studies that eventually panned out … At the start of the market, shares for every study cost $0.50 each [representing a 50–50 chance of replication]. As trading continued, those prices soared and dipped depending on the traders' activities.
>
> (Yong 2018)

After the "prediction market" was closed, "the market assigned higher odds of success for the 13 studies that were successfully replicated" (2018). Some stocks ended up with evaluations of ninety-five cents – indicating a collective confidence in replication, while others ended up at twenty-five cents – showing little confidence. The average predicated market price was $0.634 (63.4%) while the observed replication rate was an uncannily similar 61.9% (Camerer et al. 2018:639). "Peers are to some extent able to predict which studies are most likely to replicate" – presumably based on their knowledge of the original papers and the reactions of other experts betting on them (Figure 10.1).

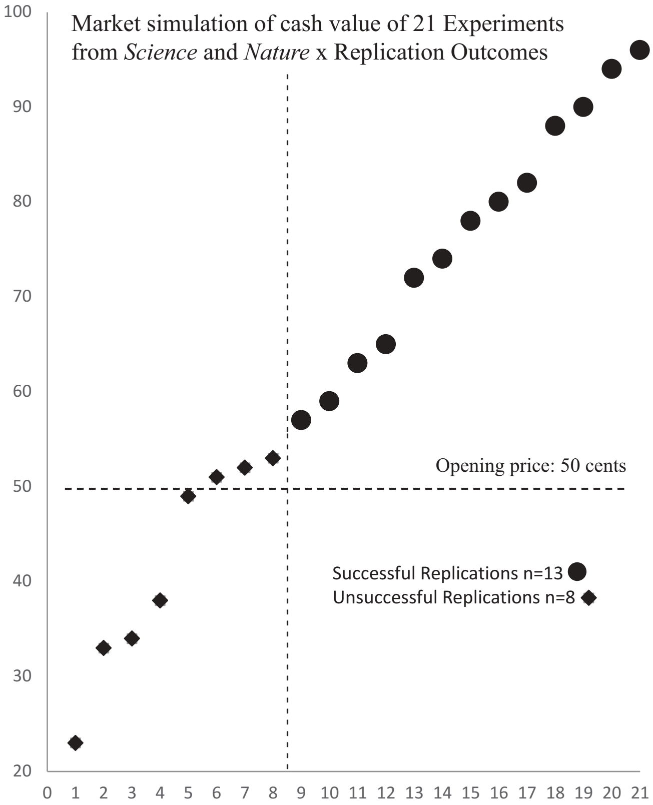

Figure 10.1

Replications in priming research

John Bargh had a moment of inspiration when he hypothesized that social behavior is often anchored in a sea of cues in the immediate environment from which our brain borrows to make sense of our choices and feelings. That is not so revolutionary. But the idea that we can absorb cues subliminally that contribute to our behaviors without a conscious endorsement or censorship of the reflective faculties of mind is a more revolutionary claim. This harkens back to the claims of "subliminal seduction" (Key 1973; Sedivy 2011). The psychologist's task was to bring subjects into the laboratory and to "prime" them directly with specific laboratory treatments, and then to determine whether different treatments had different outcomes on what appeared to be entirely unrelated situations. Bargh presented subjects with a series of words and asked them to make up sentences using the words. Other psychologists have asked subjects to unscramble a set of words to make a sentence which stimulates or primes the brain based on what the sentence conveys. One group of students was given the words: *Florida*, *forgetful*, *bald*, *gray*, *wrinkled*. These obviously primed cognitions of old age. Students were asked to compose sentences using these words. Once the students completed the experiment, they were summoned to another laboratory for a different exercise at the end of a long corridor. What Bargh and colleagues discovered was that subjects who had been primed with tokens of old age (as opposed to other concepts) took longer to walk the length of the corridor as measured by an observer with a hidden stop-watch. They were mimicking old age. This study became, in the words of Daniel Kahneman, "an instant classic" since the preoccupation with the elderly led the subjects to unconsciously mimic the elderly, however momentarily, by walking more slowly down the corridor (Kahneman 2011:53).

Researchers applied the paradigm to other areas. Could hand-washing lead to a situational change in moral judgment? Schnall, Benton, and Harvey (2008) reported that subjects who had been primed to cleanliness subsequently judged certain borderline moral actions less severely. Zhong, Strejcek, and Sivanathan (2010) reported that priming cleanliness through the use of a hand sanitizer prior to the study made subjects' moral judgments harsher. In the second case, the attribution of the purity seems to have been to oneself, thus inflating a sense of moral superiority, and in the first case the attribution is to others, making subjects less judgmental.

Attempts to replicate these studies have been mixed. And they have sometimes been accompanied by acrimony, since the failure to replicate may point, on the one side, to charges of unprofessional behavior, potential fraud, or QRPs, or, on other side, to the incompetence, envy, and bullying behavior of the replicators (Bartlett 2013). In the age of Retraction Watch (retractionwatch.com), there is no agreed protocol for how replications ought to be undertaken and evaluated. Kahneman (2014) proposed an "etiquette" under which the proposed replication would be outlined to the original authors, who would have an opportunity to assess the fairness of the replication; the entire correspondence would be transparent and on the record. Also, in an open email letter circulated widely to colleagues in social psychology after classic studies in priming failed to replicate, Kahneman warned colleagues that the replication crisis was undermining the credibility of research: "Questions have been raised about the robustness of priming results… . Your field is now the poster child for doubts about the integrity of psychological research" (cited in Schimmack, Heene, and Kesevan 2017). He went on to suggest that the situation was a looming "train wreck." Schimmack et al. (2017) suggest that the train wreck has already occurred with each wave of replication studies discussed above. "Kahneman's concerns have been largely confirmed. Major studies in social priming research have failed to replicate."

One of the most interesting replications of priming studies was published by David Shanks and colleagues from University College London. Among the many strong claims for the effects of priming is the proposition that an individual's performance on a general intelligence test can be influenced by what is on his or her mind in the period before he or she actually answered the questions. Dijksterhuis and van Knippenberg (1998) reported that individuals answered more general knowledge questions correctly after being asked to think about the attributes of a *professor* than they did after thinking about *soccer hooligans*. Their conclusions were replicated independently and extended in subsequent publications by several other research teams. This is surprising, since "decades of research has found that unconscious or subliminal influences on behaviour are exceptionally difficult to demonstrate" and if they are detected "they tend to be over extremely short time intervals (less than a second)" (Shanks et al. 2013). The studies of priming of intelligent behavior employed some sort of experimental treatment which constituted the priming, cuing, or framing of the subject, followed by the administration and completion of a written test – which would take a significant duration of time (not a few seconds). The original Dijksterhuis and van Knippenberg study measured differences between subjects randomly assigned to the *professor* or *hooligan* treatment groups. In these treatments, the subjects were asked to reflect on the attributes of the subject (*professor* on the one hand, or *soccer hooligan* on the other). They were asked to list their behaviors, lifestyles, and appearances. In a variation of the experiment, they were asked to identify their respective traits. And, in one of their replications, they asked the subjects to think, not about professors and soccer hooligans, but about the traits of *intelligence* or *stupidity*. Following these treatments, they were administered an unrelated multiple-choice general knowledge test to measure the influence of the respective primes. The professor/intelligence primes were hypothesized to enhance the knowledge score; the hooligan/stupidity primes were hypothesized to lower the scores.

In the replication study, Shanks and colleagues modified the original experiment in several ways to enhance the credibility of the replication. Instead of a short paper and pencil exercise to create priming, subjects were exposed to an eight-minute video either showing professors discussing cosmology, or a documentary on soccer hooligans. Also, subjects were given a *pre*-test on intelligence questions, as well as a *post* test, in order to determine the *change* in the intelligence measure resulting from the priming. In a second replication, the eight-minute video clip was dropped, making it more comparable to the original study. In a third replication, subjects were given a nine-minute priming procedure (which was longer than the original study) requiring subjects to imagine the traits, characteristics, etc. of the subject (*professor* or *soccer hooligan*). And, in a fourth replication, the priming exercises were limited to five minutes, but the sample size was doubled to increase statistical power. In all, Shanks and colleagues produced nine replications that reflected the existing publications that claimed to produce the enhancement of knowledge through priming subjects to images associated with higher (*professor*) or lower (*hooligan*) levels of intelligence. Moreover, in one replication, "participants were explicitly told the experimental hypothesis and the expected direction of the effect the priming manipulation might have on their performance in the general knowledge test" (Shanks et al. 2013). This was done to control for expectation effects. What was the result of this most sophisticated series of replications of startling reports of intelligence priming?

Not a single replication found evidence consistent with the original publications. They *did* find a positive effect in one test. Where subjects were given a financial incentive for performing well on the knowledge test, those subjects did better compared to subjects who did not receive an incentive, and this was independent of the non-significance of the professor/hooligan priming. In conclusion, "the results reported here suggest that priming the concept 'Professor' (versus 'Soccer hooligan') confers no advantage in answering general knowledge questions" (Shanks et al. 2013). This was the case despite the fact that they spent more time creating the priming treatment, increasing the numbers of persons in the experiments to improve the chance of a significant outcome, and even explicitly leaked the hypothesis to the subjects. This does not mean that priming "doesn't work." It means that the priming of intelligent behavior is, in Shanks's term, "an elusive phenomenon" that has not been convincingly established in credible research. Previous work may have been the subject of false positives, small sample sizes, and careless methodological practices. Shanks points out that many of the previous experiments fell just short of the significance criterion ($p < 0.05$) but were published because the results were in the predicted direction, and the phenomenon was prematurely treated as factual.

Reasons for problems in replication

There are a number of reliability problems in the literature of any existing field of empirical inquiry that can lead to difficulties in replication. *Publication bias* is the fact that journals typically do not publish negative findings. Experiments that fail to establish any significant outcomes end up in "the file drawer." The file drawer problem is that individuals typically do not know what is in their colleagues' file drawers and may undertake research that has already proven fruitless (see psychfiledrawer.org). Publication bias is especially problematic when someone undertakes a meta-analysis that pulls together all the studies of the same problem but only employs what successfully was accepted for publication (due to significance) and overlooks all the negative findings in all the file drawers. *Verification bias* is a more serious phenomenon that amounts to questionable research practices. It refers to a stubborn resistance to accepting the null hypothesis – the assumption that there is no inherent relationship between the variables being studied. The null hypothesis is the default position in experiments. This is what the researcher is attempting to eliminate through experimental investigation. For example, continuing to repeat an experiment until it "works" as desired, or excluding inconvenient cases or results may make the null hypothesis immune to the facts. Verification bias is "the use of research procedures in such a way as to 'repress' negative results by some means" (Levelt et al. 2012:47). For example, a researcher may exclude some cases because the individuals did not seem to respond to the treatment, or because they were outliers, thus reducing variance in the dependent variable and making statistical significance more likely to emerge. Or a subgroup is selected for analysis because, in retrospect, this is the group that yields the significant tests. *HARKing* – hypothesizing after the results are known – is the reconstruction of the objective of the work because some finding reaches statistical significance, even if it was not the objective of the work in the first place (Świątkowski and Dompnier 2017:114). Hence, a random event – a false positive – may be treated as an bona fide achievement and motivate others to replicate it. All these practices are what Simmons, Nelson, and Simonsohn (2011) refer to as "the researchers' degrees of freedom," meaning unjustifiable flexibility in data analysis that is undisclosed to the reader, such as employing several different measures of the dependent variable, or controlling for gender effects after the fact to determine if the effect is gender specific.

> Likewise, other procedures known as *p-hacking* (undisclosed multiple testing without adjustments) and *cherry picking* (dropping observations to reach a significance level) lead to the same problematic consequences: using such techniques when analyzing data increases the Type-1 long-term error rate, especially when applied in combination.
>
> (Świątkowski and Dompnier 2017:114)

The result is that a large portion of the published literature consists of false positives – studies whose minimum statistical acceptability has been inflated by seemingly minor adjustments to the data which pushes the test just across the 0.05 statistical threshold.

Shanks et al.'s study of priming intelligent behavior discovered that some researchers found effects for "moderately difficult questions" but not for "difficult questions" – which suggests the researchers were hacking the dependent variable by parsing it after the fact into levels of difficulty in search of a statistically significant outcome. Sometimes, researchers suppress tests of reliability in standardized measures because such tests indicate low reliability in the sample, or researchers will only report a subset of items on a scale which has proven significant where the scale as a whole is non-significant. There may be good justification for such practices, particularly in exploratory research where a scale is not well known, but these steps must be reported to readers who might attempt to replicate the research. During the investigation of Diedrick Stapel, the investigation committees were surprised by the incompleteness or inaccuracy of reporting practices in contemporary experimental research (and mentioned among other things):

- Subjects were identified vaguely as "students in the Netherlands"
- Reference was made to a well-known measurement scale but a non-standardized variation was substituted
- Where a seven-point scale was used, a five-point scale was reported
- A measurement of "attractiveness" assumed to be made by a third party was actually a self-assessment
- Researchers failed to mention that the published experiment was conducted in a session with other experimental conditions in which multiple measurements could bias one another

- The number of subjects reported was different from the number studied
- The nature and extent of missing data were unexplained

The committees concluded that "the diligent and critical handling of research and data were not held in high esteem, and were no part of the practical research education of PhD students" (Levelt et al. 2012:51–2). If their assessment is correct, this would explain the very high levels of failures to replicate that have been found in recent years and which have shaken confidence in the findings of modern experimental social psychology. However, it should be pointed out that failures to replicate are not unique to psychology. There is a grave concern that medical research, for example, which tests for the benefit of new drugs, surgical procedures, and therapies in double blind designs are also subject to unwarranted "researcher degrees of freedom," especially where enormous amounts of funds are invested in the development of these innovations, where considerable profits stand to be made and where the companies who stand to benefit are often partners in the evaluation of their utility (Sternberg 2020).

Potential remedies for replication problems

Simmons, Nelson, and Simonsohn (2020) have suggested some guidelines to be followed by researchers and reviewers to deal with the epidemic of false positive results in recent experimental social psychology. These are more than ethical guidelines and more like requirements which, if followed, would make published research more transparent and accountable. First, "authors must decide the rule of terminating data collection before data collection begins and report this rule in the article." Since making this proposal, others have suggested that the research program should be registered in a professional online repository prior to drafting the article, and should lay out the target subjects, the sample size, and the statistical methods which are to be employed in the data analysis. This curbs the temptation to arbitrarily terminate data collection when the results appear to be significant, and to engage in trial and error statistical analysis to determine which approach is significant. Second, "authors must collect at least 20 observations per cell or else provide a compelling cost-of-data collection justification." A simple 2 x 2 design would require eighty cases. Samples of less than twenty simply do not have the statistical power to detect most effects. Cells of less than twenty were typical in the priming studies replicated by Shanks et al. (2013). Third, "authors must report all variables collected in a study." They may not discover significant associations for all of them, but this prevents cherry-picking between a convenient subsample of the variables which were segregated for the purposes of reporting. Fourth, "authors must report all experimental conditions, including failed manipulations." The rationale is to prevent authors from cherry-picking outcomes that are significant and suppressing others. Where the failed manipulations are to be published is problematic, but presumably it could be on the researcher's own website. Fifth, "if observations are eliminated, authors must also report what the statistical results are if those observations are included." There may be good reasons for the exclusion of cases after the fact, but readers should be informed if this decision has altered the tests of significance, and if the exclusion appears justified. And sixth, "if an analysis includes a covariate, authors must report the statistical results of the analysis without the covariate." This makes transparent the degree to which the main effect is significant on its own, and the wisdom of including a co-variate in the first place. In the priming intelligence studies, if priming for high intelligence following the *professor* model boosts knowledge only when the subject also receives an incentive, then the effect may be due entirely to the incentive.

There are other suggestions to make the experimental methodology more robust and transparent. One is to require authors to post their raw data as summary tables on their websites so that other researchers can determine the path of inference from the information collected and the conclusions drawn in the literature. It also would permit colleagues to detect *p*-hacking and statistically suspicious results. Another suggestion is to raise the alpha level of Type-1 errors to $p < 0.01$ *or* $p < 0.001$. This may be redundant, since most researchers publish the most conservative values they detect, which may include 0.01 or smaller. In a sense, readers can already assess the likelihood of false positives under the existing rules. However, the use of the $p < 0.05$ level perhaps should be limited to exploratory studies where it would make it easier to detect novel effects, while confirmatory studies would have more stringent requirements. Another point: the Open Science Forum advocates the importance of ongoing replication practices to improve the reliability of knowledge discovered in experimental studies, and to shift professional incentives and publication opportunities for scientists who undertake these contributions. Relatedly, the PsychFileDrawer.org website has become a watchdog for questionable positive results and their investigation. And, finally, researchers should archive their data securely. The excuse that data become "lost" when colleagues have an interest in re-examining them is tantamount to their wilful suppression. None of these measures will completely reform the credibility problem associated with the recent waves of failures to replicate. They do not go far enough to curb fraud, fabrication, and plagiarism. These will not be eliminated by reducing the degrees of freedom associated with undisclosed data manipulation. And they risk going too far in the other direction by compromising the pre-registration of experimental plans. In the interests of disclosure, anyone who pre-registers a research proposal publicly risks losing his or her priority to ideas which are posted before the evidence that supports them is obtained. And that would have a detrimental effect on scientific competition.

Transforming the research process

The past few years have produced a new mindset among social psychologists. The identification of fraudulent publications was deeply disturbing to many. However, it could not have been totally surprising to those who thought that the data which were too good to be true actually turned out to be just that – untrue. But the uncovering of widespread questionable research practices among researchers who are otherwise assumed to be working in good faith was another wake-up call. And the waves of large-scale failures to replicate in prestigious specialist journals, as well as in the apex science journals, has put a large question mark over the scientific credibility of the entire field. However, the large-scale approaches to replication associated with the Open Science Forum, and the Reproducibility Project have laid the foundation for greater cooperation in primary multi-centered research that is capable of building on collective resources and sample sizes that elude individuals with limited budgets in individual research centers. The current crisis has the potential to fundamentally transform how social psychology research is conducted. In this way, a rather dark period in the history of experimental social psychology may provide the foundation for greater future advancements. However, those advancements may require us to consider the alternative prospects of a scientific life without experiments. That is the subject of the epilogue.

Epilogue: looking forward

Scientific life after experiments

Introduction: the disappearance of the social in social psychology

John Greenwood draws a startling conclusion from his history of 20th century social psychology which is captured in the title of his book: *The Disappearance of the Social in American Social Psychology* (2004a). During the first half of the 20th century, psychologists gradually moved away from a European intellectual understanding of persons which treated them primarily in terms of their group embeddedness in which the total was more than a sum of the parts. The individual was viewed as a subject shaped by a collective history and culture, as well as an agent of change in both. Specifically, in Durkheim's *Division of Labour in Society* ([1893] 1984), "the individual" was an historic development in consciousness that only became possible because of the increasing division of labour (1984:141, 238ff). Social psychology moved from this dialectical view to a radically different perspective in which groups and social aggregates were understood primarily as exogenous forces acting on, or coercing, individuals. From this new perspective, social psychology was a branch of individual psychology. Floyd Allport represented this opinion as early as 1924: "there is no psychology of groups that is not essentially and entirely a psychology of individuals" (cited in Greenwood 2004b:21). By contrast, Durkheim viewed cognitions, emotions, and behaviors as essentially collective realities that were acquired from social class, religious affiliation, and nationality. Persons were connected to one another through the division of labour which fostered interdependence. Individual agency thrived or withered through the material and cultural resources generated by society. This suggested that persons were essentially the individual components of a "common consciousness" (Durkheim 1984:233), a concept hostile to the creed of rugged individualism and rationalistic pragmatism which were dominant ideals in America's democratic tradition. Le Bon's construct of the psychology of the crowd further alienated American psychologists, since it depicted the collective mindset as irrational. This threatened the ideals of autonomy and rationality associated with liberalism, in which "individuality is conceived in terms of *independence* from social community" (Greenwood 2004b:24). Le Bon's unruly crowd acted on primitive instinct, not coordinated action. Eisenstadt's (1954) more constructive emphasis on the paramount importance of the "reference group" as an anchor of the individual's cognition, emotion, and behavior was superseded by methodologies predicated on individualism. An exception was Muzafer Sherif (1956), who reported the average attributes of groups (the "Red Devils" versus the "Bulldogs") and captured social ties in terms of "sociograms" (i.e., friendship reciprocities) in the summer camp field studies. Social psychology eventually abandoned the study of persons in their capacity as members of genuine social groups or as actors within existing reference groups in favor of persons selected without any regard to their placement in the social order.

This evolved hand in hand with an adoption of the experimental method as the *sine qua non* of scientific knowledge. Greenwood (2004b:28) writes: "when social psychologists in the 1960s abandoned any interest in exploring the social dimensions of cognition, emotion, and behavior, their commitment to methodological and statistical rigor virtually ensured that experiments would exclude any residual social dimensions." The fact that persons might have prior attachments or knowledge of one another was treated as a *contamination* of the experiment. "Any form of 'psychological connection' between participants grounded in their orientation to represented social groups constituted a source of confounding and violated standard assumptions about statistical independence" (2004b:28). Persons were to be randomly selected from the population for study, and randomly assigned to experimental and control conditions. The adoption of the experiment eliminated culture and history from social psychology and created a premium on "situationism" – the belief that the major determinants of social behavior are in the immediate environment of the actor, and that the primary unit of analysis is the individual as opposed to the higher-level reference group.

The experimental methodology was premised similarly on the elimination of "demand characteristics" (Orne 1962). Pre-existing normative assumptions that participants bring with them to the experimental setting on which they rely to interpret the situation were hounded from the laboratories as a source of pollution. Simultaneously, this shift in social psychology bereft of embedded individuals and their reference groups was accompanied by a theoretical movement to cognitive psychology. This amounted to the supposition that an explanation of the "social" in social transactions became equivalent to an explanation of how the brain "processes information." This started with cognitive dissonance and has now come to a head in the preoccupation with "priming" in modern social psychology. Festinger et al.'s (1956) analysis of cognitive dissonance assumed that some disjuncture of reality occurred at a subliminal level and resulted in an attitude change without reflection or conscious deliberation. Priming assumes the same model of the person as an automaton. As a result of the disappearance of the social from social psychology, the social dimensions, attachments, obligations, or aspirations in "the information" which the automaton processes tended to be minimized. In addition, the humanity of the participants was frequently neglected. Those who participated in the experiments were defined as "subjects," that is, organisms who were subjected to "treatments," as opposed to "volunteers" who cooperated with the scientist to advance knowledge. Volunteers in society usually have to be informed of what is expected of them and how they will be treated. Subjects apparently were recruited with little consideration of informed consent.

The difficulty of suppressing the social in the classical studies: some archival observations

Earlier in this book I explored information discovered by the "archival turn" in social psychology – the study of the documents created during the execution of some of the classical studies in experimental social psychology and collected in publicly accessible archives. The classical studies recurrently appear to have been *demonstrations* in which the participants were choreographed into roles designed to illustrate the researchers' theoretical presumptions. The archival materials frequently point to the difficulty in suppressing the social elements arising from the laboratory interaction. For example, Perry unearthed evidence that Milgram failed to debrief the plurality of his 780 participants before the experiment was ended for fear of "contaminating" future participants, and the necessity for "testing" them under conditions of social naïveté. As a consequence, many participants left the experiments under the impression that they might have injured someone. In addition, he cherry-picked results for publication and suppressed the "intimacy experiments", which recruited persons well known to one another to act as teachers and learners; these pairings led to extremely low levels of obedience. The fact that people intimate with one another could not be expected to act as automatons was inconsistent with a science which had extinguished the relevance of reference groups. He also exposed many participants to long-term traumatizing conditions which were evident in Perry's interviews of participants decades afterwards, which suggests that the designs neglected the normal sensibilities which participants take with them into the laboratory.

Perry's investigation of the Sherif archives of the Robbers Cave field studies revealed that there was no informed consent obtained from participants or their parents beforehand, and no debriefing after the young teenage boys had been used in unacknowledged field experiments for weeks at a time. Both practices were designed to ensure a guileless and passive subject body before, during, and after the studies to suppress the unwelcome intrusion of the social into social psychological experiments. The participants were treated as automatons without individual biographies. During those experiments, they were unprotected from bullying as Sherif sought to establish "natural" status hierarchies. He designed conditions explicitly to evoke mutual antagonism between his nascent groups. His pseudo-counselors engineered intergroup hostility through surreptitious damage to the tents and belongings of the participants by perpetrating vandalism against boys in one group and blaming the boys in the other, a ruse that failed and which led to the premature termination of the 1952 summer camp experiment as the boys seized control of the situation. Decades later, according to Perry, former participants resented being used in this fashion, and recalled that the counselors were not acting as expected, that is, as adults.

Le Texier's archival re-examination of the Stanford Prison Experiment revealed that the guards, contrary to what has been published about their spontaneity, were actually coached ahead of time as a group in terms of Zimbardo's expectations of their domineering behavior. The entire study was an unacknowledged plagiarism of work carried out by Zimbardo's own students months earlier. The participants assigned to the role of inmates were actually humiliated by the dress code and harassed by the guards to achieve something akin to dehumanization. Finally, Rosenhan's exposé of mental hospitals appears to have been largely fabricated through the use of non-existent "pseudo-patients," and Rosenhan himself never followed the protocols for admission for treatment. In their exalted roles as psychological experimenters, Zimbardo and Rosenhan's lack of candor and transparency showed a degree of megalomania in respect of their audiences.

However, aside from what are now deemed to be questionable research practices (QRPs), the archival materials often contain rich information about how participants actually interacted during the experiments. This was most evident in the Yale University archive for Milgram's work, which contained a treasure trove of documentation about the obedience experiments, including audiotapes of the majority of the experiments themselves. These contained the conversations between the participants and the experimenters and permitted later researchers to make an assessment of the validity of the original experiments and the explanation of behavior in terms of obedience to authority. Researchers found that what Milgram published and what actually happened were often quite different. For example, Milgram was insistent that the levels of obedience to authority were equivalent for males and females. Since the activity of obedience was essentially an act of aggression involving the administration of painful shocks to the learner, one would have assumed that females would have been less inclined to comply. This is consistent with everything known about gender differences in violence reported in criminology and found in everyday life. Milgram's published protocol for pressure on the subject to comply was a four-level verbal escalation of pressure: [Please continue; the experiment requires you to continue; it is absolutely essential that you continue; You have no other choice, you must go on]. If the participants refused to comply after the last prod, the experiment was designed to have been terminated. Perry's archival research found that, in the all-female condition, Mr. Williams would not take "no" for an answer, even after the fourth prod. The female participants were badgered to comply. Perry reports that one subject sat with a cup of coffee provided by Mr. Williams for half an hour in order to negotiate her surrender to the experiment's demands. "Williams insisted that one woman continue 26 times. He argued with two others 14 times; one, 11 times; another, nine times; another, eight times; and noted that, in the case of subject 2014, the experiment ended in an 'argument'" (Perry 2012:134). In a post-experimental debriefing with Dr. Paul Errera, women claimed they were "railroaded" into compliance (2012:135). Williams insisted relentlessly that the women comply, and Milgram got the results he wanted: obedience leading to interpersonal aggression was independent of gender. However, anyone listening to the audiotapes would draw different conclusions from those who simply read the data compiled in Milgram's tables: these suggest that men and women behave identically under pressures from authority. By today's standards, this conclusion should be retracted.

What do the archival materials tell us about obedience in Milgram's experiments?

For students of social psychology, the Milgram experiments remain a kind of goalpost for the achievements and limitations of classical social psychology. This is a result of the gravity of the topic he tackled – the Holocaust – and the immensity of the data he collected in trying to understand it. We touched briefly on the apparent intense interactional work employed by the experimenter to extract high levels of female obedience. While this outcome may have been based on the utilization of QRPs, the experiment raises three other issues. First, the audiotapes suggest that there was a sometimes rich interactional exchange between the participants and the experimenter that may shed light on their rationale for behaving as they did, or, rather, their rationale for behaving as they did *from their point of view*. This is exemplified in recent work by Hollander and Turowetz, who return the question of the social to social experimentation by exploring the role of interpersonal trust in everyday life, including occasions such as psychological experiments. They employ an approach based on ethnomethodology and conversational analysis to open up "account giving" as an alternative to the traditional causal approach behind experimentation.

Second, the examination of conversations between the participants and the experimenter raises the larger issue of the rhetorical foundations of social interaction in everyday life, and the way in which social actors employ language to advance perspectives, to persuade interlocutors to adopt certain points of view, and to justify actions and opinions. This is exemplified in Stephen Gibson's re-analysis of the conversational exchanges that are captured in his analysis of the Milgram (2013a,b) experiments. And, finally, by highlighting the conversational ebb and flow found in the Milgram experiments, and the ability of participants to challenge the rigidity of experimental protocols, we are able to examine the problem of standardization, which is taken for granted in the experimental approach, and which has plagued, as we have already seen in the previous chapter, the attempts to guarantee the replicability of experimental knowledge in social psychology.

Participant accounts and the scientific understanding of obedience and defiance

Several researchers have undertaken extensive analysis of the audio tapes which recorded the experiments (e.g. Nicholson 2011, 2015; Hoffman, Myerberg, and Morawski 2015; Russell 2018). Most recently, Hollander and Turowetz (2017) analyzed a hitherto untouched source of interviews captured on tapes in the immediate aftermath of the experiments in which participants offered accounts for their behavior. Hollander and Turowetz identified 117 recordings selected from various conditions (condition 2 – voice feedback; condition 3 – proximity; condition 20 – women as participants; condition 23 – Bridgeport offices; and condition 24 – intimate relationships). In ninety-one of these recordings, the participants gave at least one clear account or explanation of their behavior. These included forty-six cases where the participants were defined as obedient (i.e., completed all the shocks), and forty-five where the participants were defined as defiant (i.e., discontinued the experiment at some point). "In the Milgram setting, participants initially displayed trust (in Garfinkel's sense) by treating the experiment as a benign study of learning for which they had volunteered" (2017:657). However, as the tension became ratcheted up, their trust in the experiment was breached, and many stopped. However, many others normalized the situation by denying that anyone was actually experiencing harm. Hollander and Turowetz employed the detailed transcription methods developed by Harvey Sacks and Emmanuel Schegloff in the field of conversational analysis, which provide a very fine-grained reproduction of the original conversations. That enterprise is not part of the causal frame of reference that distinguishes the utilization of experiments. Milgram wanted to explain behaviors such as the mass murder of European Jews with respect to causal influences of bureaucracies on individual obedience and the agentic state which expunged autonomy. Hollander and Turowetz ask a different question: how do the participants in these experiments make sense of their own behaviors? Specifically, at the end of the experiments, what accounts do they provide of their own behaviors? And how does knowledge of such self-understandings enable the researcher to understand obedience from the point of view of the participants? One can see the element of trust operating in some of the following utterances when the participants invoke the experimenter's behavior to explain their own compliance – "if it was that serious, you would have stopped me" (p. 663) – "there must have been some reason on your part to want me to continue" (p. 664) – "there was no hesitation on your part that we continue; and if there was any question that this would affect his heart, or that he would faint, you wouldn't have allowed me to go on" (p. 666) "I knew it wouldn't hurt him … because I knew you wouldn't give it to us if it was [dangerous]" (p. 667). The subject responses are intimately embedded in the "messages" they are picking up from the experimenter's behavior. This is not abject obedience by automatons who surrender their agency under the power of bureaucracy, but compliance which is offered on the supposition that the actors can trust the experimenter to prevent them from doing any real harm.

Among those defined as "obedient," there were four recurrent accounts offered by the participants for their compliance. The most prevalent explanation was the belief that *the learner was not actually being harmed*. This was reported by thirty-three out of forty-six (72%) obedient participants (2017:660). The second most prevalent explanation was that they were simply following "the scientist's instructions" (seventeen cases). The third most prevalent explanation was that they believed in the "importance of the experiment" (eleven cases). And, finally, participants complied because of their sense of being bound by the tacit "contract" after having agreed to participate (five cases). Some participants offered more than one rationale. Among those who were defined as "defiant," three major accounts were recorded, although they overlap to some extent. First, there were twenty-five cases in which the participants were "unwilling to continue" because of risk to the learner, for religious reasons, because the experiment was against the learner's will, because the learner was suffering, and to avoid responsibility for harming the learner. The second category consisted of seventeen cases where the participants complained of the "faulty assumption" behind the experiment (i.e., that administering shocks was an effective method for improving learning). The third account (thirteen cases) resulted from the participants' reports that they were "unable to continue" because this was against the learner's wishes, due to nervousness and feeling bad about making the learner suffer. Given the assurances from Milgram that the cover story was so successful in convincing the participants in the gravity of the shocks, it is surprising that the most prevalent account from the participants themselves described in this sample was that the learner was not actually being harmed and that *this occurred among the majority of those who were obedient.* The results also show that there were *a range of factors* which were associated with compliance and defiance, that is, that there was no one process that characterized the situation.

There is other evidence from archival materials that is consistent with this surprising finding. In a questionnaire circulated to former participants following the experiment, Milgram asked participants to indicate if they fully believed that the shocks given to the learner were real or not. Research assistant Taketo Murata was then asked to compare the belief in the painfulness of the shocks and the maximum score administered by the former teachers. There were twenty-three different conditions in the series of obedience experiments. In eighteen out of twenty-three experiments, participants who "fully believed" in the reality of the shocks gave *lower* maximum shock levels than those did not fully believe in the harmfulness of the shocks – an average of 2.66 fewer shocks (Perry, Brannigan, Wanner, and Stam 2020). What this means is that not everyone was completely taken in by the cover story. Those who were more likely to accept it showed evidence of restraint compared to those who were skeptical.

Hollander and Turowetz's work also sheds some light on another emerging interpretation of the Milgram experience advanced by Haslam, Reicher Millard, and McDonald (2015). Employing responses found in the archives from participants that were gathered as part of the *post hoc* questionnaire, Haslam et al. reported that most participants "were happy to have been of service" to science by volunteering for the experiment. This was offered in support of their own explanation for obedience to authority, which is based on what they call "engaged followership." In the original experiments, the participants are torn between two parties – the participants and the scientists, and they appear to identify more with the scientist and the scientific enterprise captured by the experiment. They believe that they are "contributing to a moral, worthy, and progressive cause" (Haslam et al. 2015:60). The shocks are a necessity in this view, and even though they often may have feelings of uncertainty about the experiment, in the final analysis they "believe it to be right" (2015:78). Haslam et al. drew comments from section 13 of 1,057 transcribed comments. There were seventeen thematic sections of comments classified by Milgram. The 140 comments that Haslam et al. relied on were titled “thoughts about the value of having participated in the research.” As Perry et al. (2020:14) argue, it is unclear, however, how many *individuals* made comments that appeared in section 13, let alone whether they represented the majority of participants, as Haslam et al. claim. In addition, Milgram also included section 6, which dealt with “feelings and suspicions during study.” This included 131 entries – which is consistent with the significant levels of doubt about the reality of the shocks that arise in Hollander and Turowetz’s post-experimental interviews, and Taketo’s analysis of variations in who “fully believed” that the shocks were real. In addition, during the post-experimental interviews the participants never alluded to their engagement with science, or the experimenter’s goals: “In none of [the conversations] did participants display any orientation to *the importance of the experiment* – by sympathizing with science, identifying with E’s goals, or approving of E’s management of L’s resistance or of how the experiment studies learning” (Hollander and Turowetz 2017:666). This does not mean that the engaged followership account is untrue, but it does not appear as a dominant theme. Some people resist for religious reasons. Others, being legally trained, think they will be held to a higher moral standard than others, and are defiant. And a person trained in the military to follow orders is not in the habit of dismissing them. Many invoked the payment, which was a sort of economic obligation with a non-trivial level of compensation, as a binding consideration. Others feared legal liability. Participants were pulled by a number of attachments to prior institutions. Even though they were in an experiment, the participants’ social attachments did not take a holiday. Turowetz and Hollander argued that “‘no single social psychological process uniquely suffices to explain [participant] actions but rather that compliance resulted from multiple processes involving a complex inter-play of situational forces and individual dispositions’” (Hollander and Turowetz 2018:89). Engaged followership is a far more compelling explanation than Milgram’s “agentic state.” The agentic state was described by Milgram as a cognitive mechanism that seemed to operate subliminally to cause compliance. Engaged followership is a return of the social to social psychology, but it appears to be just one of many forces that operate in the complex environment of the obedience laboratory.

The rhetorical foundations of obedient and defiant behavior

Gibson (2019) also presents a compelling re-examination of Milgram, based on his close analysis of the audio recordings. His approach is based on a sophisticated understanding of rhetoric. We have already alluded to the four prods which Milgram used to pressure participants to overcome their reluctance to continue (“please continue,” “the experiment requires that you continue,” etc.). Gibson adds what he describes as secondary prods, such as “the shocks are painful but not damaging”; “the experimenter will be responsible for the consequences”; “you may keep the cheque even if you don’t complete the experiment,” etc. When examined in the context of conversations between participants and the scientist, these “prods” are better understood as arguments designed to compel agreement and cooperation.

From this perspective, it becomes absurd to regard the outcomes measured by Milgram "as straightforward demonstrations of people following orders" (Gibson 2019:69). Although Milgram wrote as though the prods were employed strategically according to a sequential plan which escalated the "pressure to conform," the transcripts of the conversations suggest that their use was comparatively unsystematic. Mr. Williams often deviated from the sequential protocol which was used to challenge subject resistance. As noted earlier in the all-female experiment, an escalation did not trigger the termination of the experiment as Milgram had suggested. "Milgram did not employ his experimental procedure in practice as it appears in published reports of his work" (p. 67).

Gibson's approach builds on Billig's (1996) perspective, which devotes the centerpiece in critical social psychology to the activities of rhetoric and thinking. Where Goffman's social psychology was grounded in dramaturgy and the parallels between professional acting (i.e., stage-craft) and how people present themselves to each other in everyday life, the rhetorical perspective conceptualizes persons essentially as orators. People are orators in the sense that they are immersed in ongoing conversations with others in which they engage in a play of mutual exchanges designed to influence one another. They exchange reasons for actions and offer opinions about alternative choices designed to make their perspectives persuasive. Billig uses the term "witcraft" to characterize the everyday inventory of strategies that speakers of natural language master in order to advocate situations verbally to their advantage. Gibson closely analyzed audio recordings from four key obedience experiments (experiment 2 – voice feedback; experiment 4 – touch proximity; experiment 7- two peers rebel; and experiment 20 – all female). He focused on the argumentative strategies employed by participants to justify their discontinuation in the obedience experiment. Where Hollander and Turowetz examined the accounts offered *after the fact*, Gibson captured the arguments and reasoning recorded *during* the ongoing experiments. When participants resisted pressure to comply with the scientist's expectation, they grounded their conduct in a series of compelling reasons and arguments. Participants invoked the apparent pain or danger to the learner forty-three times. They invoked the learner's withdrawal of consent thirty-nine times. In twenty-eight separate occasions, the participants offered to return the cheque as a *quid pro quo* for stopping. In fourteen occasions, they asserted their autonomy of action. In fifteen cases, they reasoned that the learner's heart condition necessitated a termination of the experiment. In fourteen cases, they argued that the learner's protests were decisive. In eleven cases, they invoked issues around who ultimately was responsible, and in ten cases, they reasoned that they could not continue knowing that they themselves would object to receiving such shocks. In all, Gibson enumerated 184 cases of recurrent arguments for resistance which were associated with defiance across four key experiments (2019:149). This approach to the experiments repopulates them with responsible actors, and repudiates the supposition of that obedience to authority is some sort of mechanical reflex.

Gibson's evidence, like that of Hollander and Turowetz, is captured in samples of transcript that show how subject resistance and compliance developed in conversations between participants and the scientist. One of Gibson's more interesting findings occurred in early experiments when participants resisted continuation until the scientist confirmed that the learner was well and was willing to continue. The scientist left the room, pretending to check on the learner and reported back that all was well, whereupon the participant continued. After consultation with Milgram, Williams subsequently resisted pressure to consult the learner. The prods were tailored to the ongoing reactions of the participants and designed to challenge arguments and reactions that might motivate the participant to desist and terminate the experiment. Condition 4 – touch proximity – presented a different scenario from the other conditions. In the other conditions, the learner's shouting and hollering were pre-recorded. Of necessity, these signs of agitation had to be replaced with the actor's performance in the same room. This permitted the subject to directly communicate with the learner and the scientist in a more spontaneous way. It was much easier for the subject to disobey after asking the learner, Mr. Wallace, if he wanted to continue. The situation also required Wallace to improvise by saying, for example, he refused to touch the plates and that he was no longer part of the experiment. Mr. Wallace's performance in this experiment was a significant deviation from the standardized version presented in the pre-recorded experiments. "By their joint resistance, the learner and the participant put up a united front against the experimenter" (Gibson 2019:142). This condition had one of the highest levels of defiance in the experimental series. The results were not a function of "physical proximity," as Milgram had labeled it, but changes in the nature of the face-to-face engagement, and the more realistic situation of interaction compared to conditions that relied exclusively on pre-recorded utterances from the learner. The other unique feature of this experiment was that the participant knew that the learner, no matter how much he protested, was present in the room, and was alive and well (2019:135). These variations make the comparison to other experimental protocols ill-advised where it was assumed officially that the only salient difference was the *proximity* between the teacher and the learner. What the transcripts reveal is that *proximity* was not a matter of physical space as much as an intensification of reciprocal social engagement between the participants, the scientist and the learner. Gibson's focus on rhetoric forces us to re-think the meaning of obedience in Milgram's work and, notwithstanding the powerful demand characteristics associated with the Yale setting, to replace a focus on "following orders" with a more negotiated transaction by participants sensitive to the contingencies of the setting. In fact, among those who were obedient, Gibson found that in conditions 2 (voice feedback) and 20 (all women), the experimenter did not have to go beyond the first prod (*please continue*) in thirty-seven out of forty-four participants. Only two participants triggered prods 3 and 4, those most likely to be characterized as orders (2019:170).

The problems of standardization in experimental social psychology

Analysis of the interaction of participants with the learner and experimenter shows how difficult it is to ensure that all the subjects are treated identically in the sense that they experience the same "treatment" before the outcome is measured. A number of researchers who have studied the audiotapes note how that the scientist often did not follow the prod strategy consistently. Gibson also found that, in several cases, the experimenter appeared to leave the laboratory to ensure that the learner was willing and able to continue. Perry observed how participants in the all-women experiment were subject to undue pressure to conform compared to other experiments. When Milgram used different actors in the new baseline experiments, the completion rate dropped from 65% with Williams and Wallace in the fifth session (cardiac condition) to 50% with Emil Elgiss and Bob Tracy in the sixth session (change in personnel), although both sets of experiments used the same protocols (Perry 2012:390). That could be a result of differences in rehearsal and/or differences in personality. Also, the experiments are radically different when the learner is present in the same room as the teacher than when he is in a different room, because this changes the opportunity for direct communication between the parties. If Milgram's case is any illustration of problems that are endemic to social psychology experiments, it should be no mystery why replications have become such a worrisome issue in the 21st century. This is not a criticism of the experiments as much as an acknowledgment of the hazards of employing sentient creatures in conditions that have any verisimilitude to everyday life. And while the designs could be much more restrictive in terms of limiting participant responses, this itself would undermine the ecological validity of the experiment. What the qualitative studies of the laboratory recommend, above all else, is the degree to which the hypothetico-deductive enterprise minimalized the potential challenge to experiments posed by the social nature of laboratory transactions.

Life after experiments: causality, accounts and understanding

It is instructive to compare the conception of explanation inherent in Milgram's endorsement of the experiment as opposed to the approach advocated by social psychologists who focus on the actual transactions which are revealed in the experiments. At its core, the experimental approach assumes causal relationships which can be described in objective language. For example, aggression against an innocent subject is caused by the coercion of persons by an authority that produces obedient outcomes. If coercive authority is required for the outcome, if the cause occurs *before* the outcome, and if the association between the cause and the outcome is not spurious, one is entitled to infer a causal effect. Understanding is achieved by identifying that effect, and that relationship is said to exist objectively. It is a different matter for Gibson and for Hollander and Turowetz. For them, understanding is achieved by determining empirically how the outcomes were negotiated and accounted for in the language and actions of the participants. However, "the facts" here are a bit more elusive, since they consist of qualitative or subjective assessments which cannot be rendered more objective by access to an underlying orderliness. For example, Hollander and Turowetz argue that the accounts offered by participants in the informal debriefings after the experiments are a valid token of what the participants thought during the experiments. They argued that the accounts *explained* the participants' choices. They were valid indicators because the memories were so fresh in the immediate period following the experiment, and the elicitation of responses was spontaneous. Critics argue that perhaps the *post hoc* accounts, particularly the claim that "no one was hurt," may have been an excuse to deflect potential recriminations for anti-social behavior. This sets up a chasm between what Gibson refers to as matters that are "under the skull" (2019:85), that is, what participants were *actually* thinking, versus what was uttered and recorded. Gibson, Blenkinsopp, Johnstone, and Marshall (2017) argue that these utterances, whether we focus on the after-the-fact debriefings or the actual experiments, should be examined only in terms of the work that they perform for the speakers in the immediate situation of their utterance. This is not to deny either that the participants really mean what they say on the one hand, or are uttering alibis on the other. "We occupy a middle ground of agnosticism in which we analyze what people say for its function in the context of which it occurs, without needing to draw speculative conclusions about whether or not it is reflective of underlying thought" (Gibson 2019:91). Importantly, in his analysis of the reasons and arguments for desistance and compliance, Gibson does not systematically report the actual associations between desistant/compliant utterances and the participants' patterns of withdrawal or obedience. That means that "understanding" phenomena of obedience and resistance has quite different implications in an experimental/causal perspective and an interpretive/discursive perspective.

The experiment epitomizes the hypothetico-deductive method. It is premised on the empirical testing of hypotheses under controlled conditions to establish regularities in social behavior akin to the regularities found in the natural sciences. From this perspective, the observer understands the phenomena if it can be explained in terms of objective statements that specify the outcomes of social forces on social conduct. By contrast, the studies of social interaction based on qualitative methods of the kinds presented in the analysis of discourse are essentially interpretive. From this perspective, the observer understands the phenomena if researchers can reliably describe the methods employed by people to make sense of their interaction through the analysis of such devices as account-giving and rhetoric. In this chapter, we have contrasted the classical perspective of Milgram, who employed a causal model of the role of authority in producing obedience, and the interpretive perspective illustrated in the work of Hollander and Turowetz and of Gibson, who explored the role of accounts and rhetoric in making the behavior of the participants in the obedience studies intelligible. On the whole, experimental social psychology has not been all that successful in achieving consensus regarding new, objective, non-trivial knowledge, despite its popularity and longevity. The classical studies suffered from what might charitably be called an abundance of verification bias, which made them more demonstrations than impartial tests of hypotheses.

Contemporary studies are marked by recurrent failures to replicate – a situation that calls into question their ability to achieve any reliable knowledge at all. Wholesale intellectual investment in the alternative discursive approaches would appear long overdue. However, there is more at stake than a quotidian decision about methods. Lewin (1931:142) suggested that it constituted a choice between an Aristotelian philosophy that is "anthropomorphic and inexact" and a Galilean philosophy which is precise and universalizing. The problems of the qualitative alternatives to experimentation are not insignificant, but move in a different direction. Take the issue of replication which we reviewed in the previous chapter. This is an issue for any study, experimental or otherwise. But it may raise a different kind of question on the qualitative side. Consider, for example, one of the greatest claims of modern interpretive social science – that Western capitalism, which emerged with such dramatic consequences in 15th century Europe, occurred as a result of changes in religious beliefs, and in the rise of ascetic Protestantism. This was a religion that encourage the accumulation of wealth as a sign of salvation. The attempts to replicate Max Weber's findings fill volumes, but the dilemma for us is relatively simple. On the one side is the proposition that the significant aspects of our culture are unique and constantly subject to change. All this can be documented, but the failure to find a similar social development elsewhere is not a sign of the failure of the Weberian thesis (Weber 1958) but an acknowledgment of the inherent idiosyncrasies of human life and history. On the other side is the supposition that history can be reduced to inherent laws which we have yet to describe (Hempel 1952, 1965). This raises the spectre of historicism raised by Popper (1960) – that history is pre-ordained by psychological and sociological laws, with the implication that human agency, freedom, and choice are illusions. This juxtaposition of historicism (determinism) and developmental relativism magnifies the dilemma of turning our backs on a deterministic approach which has yielded few returns in comparison to the natural sciences and venturing into a new land where the achievements would be palpable, but inherently subjective, that is, Aristotelian or anthropomorphic.

References

Abse, D. 1973. *The Dogs of Pavlov*. London: Valentine, Mitchell and Co.

Allport, F. H. 1924. *Social Psychology*. Boston, MA: Houghton Mifflin, (Johnson Reprint Corporation 1967).

American Association of University Women. 1992. *How Schools Shortchange Girls: The AAUW Report*: Washington, DC: Wellesley College Center for Research on Women.

American Psychological Association. 2015a. "A Reproducibility Crisis", *Monitor on Psychology 46* (9): 39. Online www.apa.org/monitor/2015/10/share-reproducibility, Accessed 14 December 2019.

American Psychological Association. 2015b. "APA Review Confirms Link between Playing Violent Video Games and Aggression – Finds Insufficient Research to Link Violent Game Play to Criminal Violence," Press Release online www.apa.org/news/press/releases/2015/08/violent-video-games. Accessed 31 January 2020.

Andersen, C. A., A. Shibuya, N. Ihori, E. L. Swing, B. J. Bushman, A. Sakamoto, H. R. Rothstein and M. Saleem 2010. "Violent Video Game Effects on Aggression, Empathy, and Prosocial Behavior in Eastern and Western Countries", *Psychological Bulletin 136*: 151–173.

Anderson, D. F. and R. Rosenthal 1968. "Some Effects of Interpersonal Expectancy and Social Interaction on Institutionalised Retarded Children", *Proceedings of the 76th Annual Convention of the American Psychological Association 3*: 479–480.

Archer, J. 2000. "Sex Differences in Physical Aggression to Partners", *Psychological Bulletin 126* (5): 697–702.

Arendt, H. 1964. *Eichmann in Jerusalem: A Report on the Banality of Evil*. New York: Penguin Books.

Aronson, E. 1999. "Adventures in Experimental Social Psychology: Roots, Branches, and Sticky New Leaves". Pp. 82–113 in *Reflections on 100 Years of Experimental Social Psychology*, edited by A. Rodrigues and R. Levine. New York: Basic Books.

Asch, S. 1951. "Effects of Group Pressure upon the Modification and Distortion of Judgements". Pp. 177–190 in *Groups, Leadership and Men*, edited by H. Guetzkow. Pittsburgh, PA: Carnegie Press.

Asch, S. 1952. *Social Psychology*. Englewood Cliffs, NJ: Prentice-Hall.

Asch, S. 1955. "Opinions and Social Pressure", *Scientific American 193* (5, November): 31–35.

Asch, S. 1956. "Studies of Independence and Conformity: A Minority of One against A Unanimous Majority", *Psychological Monographs: General and Applied 70* (146): 1–70.

Asch, S. 1958. "Review of *A Theory of Cognitive Dissonance* (Leon Festinger)", *Contemporary Psychology 3*: 194–195.

Aschwanden, C. and M. Koerth 2016. "How Two Grad Students Uncovered an Apparent Fraud", *FiveThirtyEight*, online https://fivethirtyeight.com/features/how-two-grad-students-uncovered-michael-lacour-fraud-and-a-way-to-change-opinions-on-transgender-rights/accessed 22 February 2020.

Baker, J. P. and J. L. Crist 1971. "Teacher Expectancies: A Review of the Literature". Pp. 48–64 in *Pygmalion Reconsidered*, edited by Janet D. Elashoff and Richard E. Snow. Worthington, OH: Jones.

Baker, Peter C. 2014. "Missing the Story: How Turning the Murder of Kitty Genovese into a Parable Erased Its Particulars," *The Nation* on-line. Accessed 23 December 2019.

Bandura, A., D. Ross and S. A. Ross 1963. "Imitation of Film-Mediated Aggressive Models", *Journal of Abnormal and Social Psychology 66*: 3–11.

Bandura, A. 1973. *Aggression: A Social Learning Analysis*. Englewood Cliffs, NJ: Prentice Hall.

Barker, M. 1984. *A Haunt of Fears: The Strange History of the British Horror Comics Campaign*. London: Pluto Books.

Barker, M. 1989. *Comics: Ideology, Power and the Critics*. Manchester: Manchester University Press.

Bartlett, T. 2013 "Power of Suggestion," *Chronicle of Higher Education*, January online https://www.chronicle.com/article/Power-of-Suggestion/136907

Baumrind, D. 1964. "Some Thoughts on Ethics of Research: After Reading Milgram's 'Behavioral Study of Obedience.'", *American Psychologist 19* (2): 421–423.

Baumrind, D. 1985. "Research Using Intentional Deception: Ethical Issues Revisited", *American Psychologist 40* (2): 165–174.

Bem, D. 2011. "Feeling the Future: Experimental Evidence for Anomalous Retroactive Influence on Cognition and Affect", *Journal of Personality and Social Psychology 100* (3): 407–425. DOI: 10.1037/a0021524.

Bentham, J. [1789] 1970. *An Introduction to the Principles of Morals and Legislation*. London: Athlone.

Berkowitz, L. 1971. "Sex and Violence: We Can't Have It Both Ways", *Psychology Today 5*: 14–23.

Berkowitz, L. 1999. "On the Changes in U.S. Social Psychology: Some Speculations". Pp. 158–169 in *Reflections on 100 Years of Experimental Social Psychology*, edited by A. Rodrigues and R. Levine. New York: Basic Books.

Berkowitz, L. and E. Donnerstein 1982. "External Validity Is More Than Skin Deep", *American Psychologist 37* (3): 245–257.

Bhattacharjee, Y. 2013. "The Mind of a Con Man," The New York Times Magazine April 26, online, www.nytimes.com/2013/04/28/magazine/diederik-stapels-audacious-academic-fraud.html, accessed 21 February 2020.

Billig, M. 1996. *Arguing and Thinking: A Rhetorical Approach to Social Psychology*, 2/e, Cambridge: Cambridge University Press.

Blass, T. 2004. *The Man Who Shocked the World: The Life and Legacy of Stanley Milgram*. New York: Basic Books.

Blass, T. 1992. "The Social Psychology of Stanley Milgram", *Advances in Experimental Social Psychology 28*: 277–329.

Blum, Milton L. 1949. *Industrial Psychology and Its Social Foundations*. New York: Harper.

Boyd, Edwin J. and A. Brannigan 1991. “Attitudes Towards Sexual Aggression, Robbery and Altruism: Contributions of Media Exposure versus Sociological Variables”, *Aggressive Behavior 17* (2): 61–62.

BPA. 2011. “Questionable Research Practices are Rife in Psychology, Survey Suggests”, *British Psychological Association Research Digest*, 1 December, online http://bps-research-digest.blogspot.com/2011/12/questionable-research-practices-are.html, accessed 23 February 2020.

Brabeck, M. 1983. “Moral Judgement: Theory and Research on Differences between Males and Females”, *Developmental Review 3* (3): 274–291.

Brannigan, A. 2013a. “Stanley Milgram’s Obedience Experiments: A Report Card 50 Years Later”, *Society: Transactions and Modern Social Science 50* (6): 623–628.

Brannigan, A. 2013b. *Beyond the Banality of Evil: Criminology and Genocide*. Oxford, UK: OUP.

Breland, K. and M. Breland 1966. *Animal Behavior*. New York: Macmillan.

Brookman, D. and J. Kalla 2015. “Irregularities in LaCour (2014)” online http://web.stanford.edu/~dbroock/broockman_kalla_aronow_lg_irregularities.pdf, accessed 21 February 2020.

Brown, N.J.L. 2014 *Faking Science*, Translation of Ontsporing by D. Stapel, online at http://nick.brown.free.fr/stapel

Brown, Nicholas J.L., A. Sokal and H. Friedman 2013. “The Complex Dynamics of Wishful Thinking: The Critical Positivity Ratio”, *American Psychologist 68* (9): 801–833. DOI: 10.1037/a0032850.

Brown v. Entertainment Merchants Association. 2011. *131 S. Ct. 2729*. Retrieved from www.supremecourt.gov/opinions/10pdf/08 – 1448.pdf.

Browning, Christopher R. 1998. *Ordinary Men: Reserve Police Battalion 101 and the Final Solution in Poland. With New Afterword*. New York: Harper Collins Books.

Burger, J. 2009. “Replicating Milgram: Would People Still Obey Today?”, *American Psychologist 64* (1): 1–11.

Burger, J., Z. M. Girgis and C. C. Manning 2011. “In Their Own Words: Explaining Obedience to Authority through an Examination of Participants’ Comments”, *Social Psychological and Personality Science 2*: 460–466.

Buss, D. 1994. *The Evolution of Desire: Strategies of Human Ratings*. New York: Basic Books.

Buss, D. 1995. “Evolutionary Psychology: A New Paradigm for Psychological Science”, *Psychological Inquiry 6* (1): 1–30.

Cahalan, S. 2019. *The Great Pretender: The Undercover Mission that Changed Our Understanding of Madness*. New York: Grand Central Publishing.

Camerer, C.F. et al. 2018. “Evaluating the Replicability of Social Science Experiments in *Nature* and *Science* between 2010 and 2015”, *Nature Hum Behaviour 2*: 637–644. DOI: https://doi.org/10.1038/s41562-018-0399-z.

Campbell, Donald T. and Julian C. Stanley 1963. *Experimental and Quasi-Experimental Designs for Research*. Chicago, IL: McNally.

Canada. 2019. *The National Inquiry into Missing and Murdered Indigenous Women and Girls*, Ottawa: www.mmiwg-ffada.ca/final-report/. For the Executive Summary see www.mmiwg-ffada.ca/wp-content/uploads/2019/06/Executive_Summary.pdf, Accessed 3 June 2019.

Carey, Alex R. 1967. “The Hawthorne Studies: A Radical Criticism”, *American Sociological Review 32*: 403–416.

Cattell, R. 1988. “Psychological Theory and Scientific Method”. Pp. 3–20 in *Handbook of Multivariate Experimental Psychology*, edited by R. Cattell. Chicago, IL: Rand McNally.

CBC. 2011. *Gender Gap in Tertiary Education in Canada*, Ottawa: Conference Board of Canada, www.conferenceboard.ca/hcp/Details/education/gender-gap-tertiary.aspx?AspxAutoDetectCookieSupport=1. Accessed 13 February 2020.

Centerwall, Brandon S. 1993. "Television and Violent Crime", *Public Interest* Spring 111: 56–71.

Chaffee, S. H., G. Gerbner, B. A. Hamburg, C. M. Pierce, E. A. Rubinstein, A. E. Siegel and J. L. Singer. 1984. "Defending the Indefensible", *Society* September/October 21(6): 30–35.

Chapanis, N. P. and A. Chapanis 1964. "Cognitive Dissonance: Five Years Later", *Psychological Bulletin 61* (1): 1–22.

Chase, S. 1941. "What Makes the Worker like to Work?", *Reader's DigestFebruary 38*: 15–20.

Cherry, F. 1995. *The Stubborn Particulars of Social Psychology*. London: Routledge.

Chiappe, D. and R. Gardner 2012. "The Modularity Debate in Evolutionary Psychology", *Theory and Psychology 22* (5): 669–682.

Chodorow, N. 1978. *The Reproduction of Mothering: Psychoanalysis and the Sociology of Gender*. Berkeley, CA: University of California Press.

Cicourel, A. 1964. *Method and Measurement in Sociology*. New York: Free Press.

Clark, K. B. 1963. *Prejudice and Your Child*. Boston, MA: Beacon.

Cohen, S. and E. Nagel 1934. *An Introduction to Logic and Scientific Method*. New York: Harcourt.

Colby, A. and W. Damon 1983. "Listening to A Different Voice: A Review of Gilligan's In A Different Voice", *Merrill-Palmer Quarterly 29* (4): 473–481.

Conn, L. K., C. N. Edwards, R. Rosenthal and D. Crowne 1968. "Perception of Emotion and Response to Teacher's Expectancy by Elementary School Children", *Psychological Reports 22*: 27–34.

Cook, Karen S. and T. Yamagishi 2008. "A Defense of Deception on Scientific Grounds", *Social Psychology Quarterly 71* (3): 215–212.

Cook, K. 2014. *The Murder, the Bystanders, the Crime that Changed America*. New York: Norton.

Cooper, J. and R. H. Fazio 1984. "A New Look at Dissonance Theory", *Advances in Experimental Social Psychology 17*: 229–266.

Cooper, R. M. 1982. "The Passing of Psychology", *Canadian Psychology 24* (2): 264–267.

Cosmides, L., and J. Tooby, 1992. "Cognitive adaptations for social exchange" in *The adapted mind*, edited by J. Barkow, L. Cosmides, and J. Tooby (Eds.). New York: Oxford University Press.

Crist, J. 1948. "Horror in the Nursery", *Collier's Magazine* (March*27*): 22–23, 95–97).

Cronbach, L. J. 1975. "Five Decades of Public Controversy over Mental Testing", *American Psychologist 30*: 1–14.

Daly, M. and M. Wilson 1988. *Homicide*. Hawthorne, NY: Aldine de Gruyter.

Darley, J. D. and Bibbe Latané 1968. "Bystander Intervention in Emergencies: Diffusion of Responsibility", *Journal of Personality and Social Psychology 8*: 377–383.

Davis, M. (1971) "Towards a phenomenology of sociology and a sociology of phenomenology", *Philosophy of the Social Sciences 1*(4): 309–344.

Dawkins, R. [1976] 1990. *The Selfish Gene*. Oxford: Oxford University Press.

De Boer, H., A. C. Timmermans and M. P. C. van der Werf 2018. "The Effect of Teacher Expectations Interventions on Teachers' Expectations and Student Achievement: Narrative Review and Meta-analysis", *Educational Research and Evaluation 24* (3–5): 180–200.

De Grazia, E. 1992. *Girls Lean Back Everywhere: The Law of Obscenity and the Assault on Genius*. New York: Random House.

De May, J. 2006. "Kitty Genovese: The Popular Account in Mostly Wrong". Retrieved June 30, 2007, from www.oldkewgardens.com/ss-nytimes-3.html. Also see https://web.archive.org/web/20090330005400/http://kewgardenshistory.com/kitty_genovese.html

Dennett, D. 1995. *Darwin's Dangerous Idea: Evolution and the Meanings of Life*. New York: Simon and Schuster.

Desmond, A. J. 1975. *The Hot-Blooded Dinosaurs*. London: Futura.

Deutsch, M. 1999. "A Personal Perspective on the Development of Social Psychology in the Twentieth Century". Pp. 1–34 in *Reflections on 100 Years of Experimental Social Psychology*, edited by A. Rodrigues and R. Levine. New York: Basic Books.

Dewey, J. 1901. *Psychology and Social Practice*. Chicago, IL: University of Chicago Press.

Dewey, J. [1922] 1950. *Human Nature and Conduct: An Introduction to Social Psychology.* New York: Modern Library.

Diamond, M. and A. Uchiyama 1999. "Pornography, Rape, and Sex Crimes in Japan", *International Journal of Law and Psychiatry 22* (1): 1–22.

Dijksterhuis A. and A. van Knippenberg (1998). "The relation between perception and behavior, or how to win a game of Trivial Pursuit," *Journal of Personality and Social Psychology* 74: 865–877.

Donnerstein, E. 1980. "Aggressive Erotica and Violence against Women", *Journal of Personality and Social Psychology 39*: 269–277.

Donnerstein, E. 1983. "Erotica and Human Aggression". Pp. 53–81 in *Aggression: Theoretical and Empirical Reviews*, edited by R. Geen and E. Donnerstein. New York: Academic Press.

Dunlap, David W. 2016. "1964 How Many Witnessed the Murder of Kitty Genovese?" *The New York Times*, April 6. Accessed online 18 December 2019.

Dupré, J. 2003. *Human Nature and the Limits of Science*. Oxford: Oxford University Press.

Durkheim, E. [1893] 1984. *The Division of Labour in Society.* Translated from the French by W.D. Halls. New York: The Free Press.

Dworkin, A. 1985. "Against the Male Flood: Censorship, Pornography and Equality", *Harvard Women's Law Journal 8* (Spring): 1–29.

Economist. 2013. "Trouble at the Lab: Scientists like to Think of Science as Self-correcting. To an Alarming Degree It Is Not", October 19th, online www.economist.com/briefing/2013/10/18/trouble-at-the-lab, accessed 15 January 2020.

Economist. 2015. "Try Again: A Large Study Replicates Psychology Experiments", August 27th, online www.economist.com/science-and-technology/2015/08/27/try-again, accessed 15 January 2020.

Editorial. 2017. "A New Look at the Killing of Kitty Genovese: The Science of False Confessions," Association for Psychological Science on-line. www.psychologicalscience.org/publications/observer/obsonline/a-new-look-at-the-killing-of-kitty-genovese-the-science-of-false-confessions.html.

Eisenberg, N. and R. Lennon 1983. "Sex Differences in Empathy and Related Capacities", *Psychological Bulletin 84* (4): 712–722.

Eisenstadt, S.M. 1954. "Reference Group Behavior and Social Integration", *American Sociological Review 19* (2): 175–185.

Elashoff, J. D. and R. E. Snow 1971. *Pygmalion Reconsidered*. Worthington, OH: Jones.

Elflein, J. 2019. *Death Rate for Suicide in the U.S. 1950-2017, by Gender*, www.statista.com/statistics/187478/death-rate-from-suicide-in-the-us-by-gender-sinc, accessed 13 Febraury 2020.

Ellison, K. 2015. “Being Honest about the Pygmalion Effect,” Discover Magazine, October 28. Online www.discovermagazine.com/mind/being-honest-about-the-pygmalion-effect. Retrieved 21 January 2020.

Elms, A. C. 1975. “The Crisis of Confidence in Social Psychology”, *American Psychologist* October *30* (10): 967–976.

Eron, L. 1987. “The Development of Aggressive Behavior from the Perspective of a Developing Behaviorism”, *American Psychologist 42*: 435–442.

Evans, J. T. and R. Rosenthal 1969. “Interpersonal Self-Fulfilling Prophecies: Further Extrapolation from the Laboratory to the Classroom”, *Proceedings of the 77th Annual Convention of the American Psychological Association 4*: 371–372.

Fenigstein, A. 2015. “Milgram’s Shock Experiments and the Nazi Perpetrators: A Contrarian Perspective on the Role of Obedience Pressures during the Holocaust”, *Theory and Psychology 25* (5): 581–598.

Ferguson, Christopher J. 2013. “Violent Video Games and the Supreme Court”, *American Psychologist 68* (2): 57–74.

Ferguson, C. J. 2016. “New Evidence Suggests Media Violent Effects May Be Minimal,” *Psychiatric Times 33* (11) November 24. Online. www.psychiatrictimes.com/trauma-and-violence/new-evidence-suggests-media-violence-effects-may-be-minimal. Accessed 31 January 2020.

Ferguson, C. J. and J. Kilburn 2010. “Much Ado about Nothing: The Misestimation and Overestimation of Video Violent Game Effects on in Eastern and Western Nations: Comment on Anderson et al”, *Psychological Bulletin 136* (2): 174–178.

Ferguson, K. G. 1983. “Forty Years of Useless Research?”, *Canadian Psychology 24* (2): 153–204.

Feshbach, S. and R. D. Singer 1971. *Television and Aggression: An Experimental Field Study*. San Francisco, CA: Jossey-Bass.

Festinger, L. 1954. “Laboratory Experiment”. Pp. 136–154 in *Research Methods in the Behavioral Sciences*, edited by L. Festinger and D. Katz. London: Staples.

Festinger, L. 1980. “Looking Backward”. Pp. 236–254 in *Retrospections on Social Psychology*, edited by L. Festinger. New York: Oxford University Press.

Festinger, L. 1987. “A Personal Memory”. Pp. 1–9 in *A Distinctive Approach to Psychological Research: The Influence of Stanley Schachter*, edited by N. E. Grunberg, R. E. Nisbett, J. Rodin and J. E. Singer. Hillsdale, NJ: Lawrence Erlbaum Associates.

Festinger, L. and J. Carlsmith 1959. “Cognitive Consequences of Forced Compliance”, *Journal of Abnormal and Social Psychology 58*: 203–210.

Festinger, L., H. W. Riecken and S. Schachter 1956. *When Prophecy Fails*. Minneapolis: University of Minnesota Press.

Festinger, L., S. Schachter and K. Back 1950. *Social Pressures in Informal Groups*. New York: Harper and Row.

Fischer, P., T. Greitemeyer, F. Pollozek and D. Frey 2006. “The Unresponsive Bystander: Are Bystanders More Responsive in Dangerous Emergencies?” *European Journal of Social Psychology 36*: 267–278. DOI: 10.1002/ejsp.297.

Fischer, P., J. I. Krueger, T. Greitemeyer, C. Vogrincic, M. Heene, M. Wicher, M. Kainbacher, A. Kastenmüller and D. Frey 2011. “The Bystander-Effect: A Meta-analytic Review on Bystander Intervention in Dangerous and Non-dangerous Emergencies”, *Psychological Bulletin 137* (4): 517–537.

Fisher, W. and A. Barak 1991. “Pornography, Erotica, and Behavior: More Questions than Answers”, *International Journal of Law and Psychiatry 14*: 65–83.

Fisher, W. and G. Grenier 1994. "Violent Pornography, Antiwoman Thoughts and Antiwomen Acts: In Search of Reliable Effects", *Journal of Sex Research 31* (1): 23–38.

Fowles, J. 1999. *The Case for Television Violence*. London: Sage.

Frankford-Nachmias, C. 1999. *Social Statistics for a Diverse Society.* Thousand Oaks, CA: Pine Forge.

Fredrickson, B. L. and M. F. Losada 2005. "Positive Affect and the Complex Dynamics of Human Flourishing", *American Psychologist 60*: 678–686. DOI: 10.1037/0003-066X.60.7.678.

Freedman, J. L. 1984. "Effects of Television Violence on Aggression", *Psychological Bulletin 96*: 227–246.

Freedman, J. L. 1986. "Television Violence and Aggression: A Rejoinder", *Psychological Bulletin 100*: 372–373.

Freedman, J. L. 1988. "Television and Aggression: What the Research Shows". Pp. 144–162 in *Television as a Social Issue*, edited by S. Oskamp. Newbury Park, CA: Sage.

Freud, S. [1957] 2001. *Leonardo DaVinci: A Memoir of His Childhood*. Translated by Alan Dyson. London and New York: Routledge Classics.

Friedman, H. L. and J. L. Nicholas Brown 2018. "Implications of Debunking the 'Critical Positivity Ratio" for Humanistic Psychology", *Journal of Humanistic Psychology 58* (3): 239–261. DOI: 10. 1 177/0022167818762227.

Friedman, W. J., A. B. Robinson and B. L. Friedman 1987. "Sex Differences in Moral Judgments?" *Psychology of Women Quarterly ll* (1): 37–46.

Gadlin, H. and G. Ingle 1975. "Through the One-Way Mirror: The Limits of Experimental Self-Reflection", *American Psychologist* 30: 1003–1009.

Gadow, K. D. and J. Sprafkin 1989. "Field Experiments of Television Violence with Children: Evidence for an Environmental Hazard?", *Pediatrics 83* (3): 399–405.

Gansberg, M. 1964. "37 who saw murder didn't call the police," *New York Times*, March 27, p. 1.

Garfinkel, H. 1967. *Studies in Ethnomethodology.* Englewood Cliffs, NJ: Prentice Hall.

Gauntlett, D. 1995. *Moving Experiences: Understanding Television's Influences and Effects*. London: J. Libbey.

Gerard, H. 1999. "A Social Psychologist Examines His past and Looks to the Future". Pp. 47–81 in *Reflections on 100 Years of Experimental Social Psychology*, edited by A. Rodrigues and R. Levine. New York: Basic Books.

Gibbs, J. C., K. D. Arnold and J. E. Burkhart 1984. "Sex Differences in the Expression of Moral Judgment", *Child Development 55*: 1040–1043.

Gibson, S. 2013a. "The Last Possible Resort': A Forgotten Prod and the in Situ Standardization of Stanley Milgram's Voice-feedback Condition", *History of Psychology 16*: 177–194. DOI: doi/epdf/10.1111/bjso.12272.

Gibson, S. 2013b. "Milgram's Obedience Experiments: A Rhetorical Analysis", *British Journal of Social Psychology 52*: 290–309.

Gibson, S., G. Blenkinsopp, L. Johnstone, and A. Marshall 2017. "Just Following Orders? The Rhetorical Invocation of 'Obedience' in Stanley Mil- gram's Post-experiment Interviews", *European Journal of Social Psychology.* doi:10.1002/ejsp.2351.

Gibson, S. 2019. *Arguing, Obeying and Defying: A Rhetorical Perspective on Stanley Milgram's Obedience Experiments*, Cambridge UK: Cambridge University Press, 2019.

Gillespie, R. 1991. *Manufacturing Knowledge: A History of the Hawthorne Experiments*. Cambridge, MA: Cambridge University Press.

Gilligan, C. [1982] 1993. *In a Different Voice: Psychological Theory and Women's Development*. Cambridge, MA: Harvard University Press.

Ginsberg, M. 1942. *The Psychology of Society*, (5th ed.). London: Methuen.

Glueck, S. and E. Glueck 1950. *Unraveling Juvenile Delinquency.* Cambridge, MA: Harvard University Press.

Goffman, E. 1961. *Asylums*. Chicago, IL: Aldine.

Goldhagen, D. 1997. *Hitler's Willing Executioners: Ordinary Germans and the Holocaust*. New York: Random House.

Gottfredson, M. R. and T. Hirschi 1990. *A General Theory of Crime*. Stanford, CA: Stanford University Press.

Goudge, T. A. 1961. *The Ascent of Life: A Philosophical Investigation of the Theory of Evolution*. Toronto, CA: University of Toronto Press.

Gould, S. J. 1978. "Sociobiology: The Art of Story-Telling", *New Scientist 16*: 530–533.

Gould, Stephen J. 1989. *Wonderful Life: The Burgess Shale and the Nature of History.* New York: Norton.

Greeno, C. G. and E. E. Maccoby 1986. "How Different Is the 'Different Voice'?", *Signs 11*: 310–316.

Greenwood, J. D. 2004a. *The Disappearance of the Social in American Social Psychology.* Cambridge, MA: Cambridge University Press.

Greenwood, J. D. 2004b. "What Happened to the 'Social' in Social Psychology?", *Journal for the Theory of Social Behavior 34* (1): 19–34.

Griggs, R. A. 2015a. "The Disappearance of Independence in Textbook Coverage of Asch's Social Pressure Experiments", *Teaching of Psychology 42* (2): 137–142.

Griggs, R. A. 2015b. "The Kitty Genovese Story in Introductory Psychology Textbooks: Fifty Years Later", *Teaching of Psychology 42* (2): 149–152. DOI: 10.1117/0098628315573138.

Griggs, R.A. and G.I. Whitehead 2015. "Coverage of Recent Criticisms of Milgram's Obedience Experiments in Introductory Social Psychology Textbooks", *Theory and Psychology 25* (5): 564–580.

Grossi, G. 2014. "A Module Is A Module Is A Module: Evolution of Modularity in Evolutionary Psychology", *Dialectical Anthropology 38*: 333–351.

Grossi, G., S. Kelly, A. Nash and G. Parameswaran 2014. "Challenging Dangerous Ideas: A Multi-disciplinary Critique of Evolutionary Psychology", *Dialectical Anthropology 38*: 281–285.

Haberman, C. 2016. "Remembering Kitty Genovese" Retro Report, *The New York Times* April 10. Accessed online on 18 December 2019.

Haney, C., W. C. Banks and P. Zimbardo 1973. "Interpersonal Dynamics in a Simulated Prison", *International Journal of Criminal Penology 1*: 69–97.

Haney, C. and P. Zimbardo 1977. "The Socialization into Criminality: On Becoming a Prisoner and a Guard". Pp. 198–223 in *Law Justice and the Individual in Society: Psychological and Legal Issues*, edited by J. Levine. New York: Holt, Reinhart and Winston.

Harari, H., O. Harari and R. V. White 1985. "The Reaction to Rape by American Male Bystanders", *The Journal of Social Psychology 125* (5): 653–658.

Harcourt, A. H., P. H. Harvey, S. G. Larson and R. V. Short 1981. "Testis Weight, Body Weight and Breeding Systems in Primates", *Nature* (3 September) *293*: 55–57.

Haslam, S. A., S. D. Reicher, K. Millard and R. McDonald. 2015. "'Happy to Have Been of Service': The Yale Archive as a Window into the Engaged Followership of Participants in Milgram's 'Obedience' Experiments'", *British Journal of Social Psychology 54*:55–83. doi:10.1111/bjso.12074.

Hattie, J. A. C. 2009. *Visible Learning: A Synthesis of over 800 Meta-analyses Relating to Achievement*. London: Routledge.

Hempel, C. G. 1952. *Fundamentals of Concept Formation in Empirical Science*. Chicago, IL: University of Chicago Press.

Hempel, C. G. 1965. *Aspects of Scientific Explanation*, New York: Macmillan.

Hendricks, M. 2000. "Into the Hands of Babes", *Johns Hopkins Magazine 52* (4): 12–17. on-line, September).

Henshel, R. L. 1980. "The Purpose of Laboratory Experiments and the Virtues of Deliberate Artificiality", *Journal of Experimental Social Psychology 16*: 416–478.

HEPI. 2016. *Boys to Men: The Underachievement of Young Men in Higher Education*, Report 84 (Nick Hillman and Nicholas Robinson), Oxford UK: Higher Education Policy Institute. www.hepi.ac.uk/wp-content/uploads/2016/05/Boys-to-Men.pdf, accessed 13 February 2020.

Herrnstein, R. J. and C. Murray 1994. *The Bell Curve: Intelligence and Class Structure in American Life*. New York: Free Press.

Hewstone, M., A. S. R. Manstead and W. Stroebe (Eds.). 1997. *The Blackwell Reader in Social Psychology*. Oxford: Blackwell.

Hilgard, J., C. R. Engelhardt and J. N. Rouder 2017. "Overstated Evidence for Short-term Effects of Violent Games on Affects and Behavior: A Re-analysis of Andersen et al (2010)", *Psychological Bulletin 143* (7): 757–774.

Hobson v. Hansen, 269 F. Supp. 401 (D.D.C. 1967) US District Court for the District of Columbia - 269 F. Supp. 401 (D.D.C. 1967) June 19, 1967.

Hoffman, E., N. R. Myerberg and J. G. Morawski 2015. "Acting Otherwise: Resistance, Agency and Subjectivities in Milgram's Studies of Obedience", *Theory and Psychology 25* (5): 670–689. DOI: 10.1 1777/0959354315608705.

Hoffman, M. L. 1977. "Sex Differences in Empathy and Related Behaviors", *Psychological Bulletin 84* (4): 712–722.

Hollander, Matthew H. and Jason Turowetz. 2017. "Normalizing Trust: Participants' Immediately Posthoc Explanations of Behaviour in Milgram's 'Obedience' Experiments", *British Journal of Social Psychology 56*:655–74. 10.1111/ bjso.12206.

Hollander, M. H. and J. Turowetz 2018. "Multiple Compliant Processes: A Reply to Haslam and Reicher on the Engaged Followership Explanation of 'Obedience' in Milgram's Experiments", *British Journal of Social Psychology 57*:301–309. doi:10.1111/ bjso.12252.

Hopkins, P. 1938. *The Psychology of Social Movements: A Psychoanalytic View of Society*. London: Allen & Unwin.

Houston, J. 1983. "Psychology: A Closed System of Self-Evident Information?", *Psychological Reports 52*: 203–208.

Hovland, C. 1959. "Reconciling Conflicting Results Derived from Experimental and Survey Studies of Attitude Change", *American Psychologist 14*: 8–17.

Huesmann, L. R., L. D. Eron, M. M. Lefkowitz and L. O. Walden 1984. "Stability of Aggression over Time and Generations", *Developmental Psychology 20*: 1120–1134.

Huesmann, L. R., L. D. Eron, M. M. Lefkowitz and L. O. Walder 1973. "Television Violence and Aggression: The Causal Effect Remains", *American Psychologist 28*: 617–620.

Hunt, M. 1993. *The Story of Psychology*. New York: Doubleday.

Jaffe, D. (1971). "A simulated prison" Term Paper, Stanford University Digital Repository, Philip g. Zimbardo Papers, https://purl.stanford.edu/cj735mr4214.

Jaffee, S. and J. S. Hyde 2000. "Gender Differences in Moral Orientation: A Meta-analysis", *Psychological Bulletin 126* (5): 703–726.

Jarrett, C. 2008. "Foundations of Sand", *The Psychologist 21*: 756–750. Online. https://thepsychologist.bps.org.uk/volume-21/edition-9/foundations-sand#. Accessed 25 January 2020.

John, L. K., G. Loewenstein and D. Prelec 2012. "Measuring the Prevalence of Questionable Research Practices with Incentives for Truth Telling", *Psychological Science 23* (5): 524–532. www.jstor.org/stable/41489734.

Jones, S. 1992. "Was There a Hawthorne Effect?", *American Journal of Sociology 98* (3): 451–468.

Kahneman, D. 2011. *Thinking Fast and Slow*. New York: Random House.

Kahneman, D. 2014. "A New Etiquette for Replication", *Social Psychology 45* (4): 299–311.

Kaplan, R. M. and R. D. Singer 1976. "Television Violence and Viewer Aggression: A Re-Examination of the Evidence", *Journal of Criminal Psychopathology 3*: 112–137.

Karpf, F. B. 1932. *American Social Psychology: Its Origins, Development and European Background*. New York: McGraw-Hill.

Kassin, S. M. 2017. "The Killing of Kitty Genovese: What Else Does This Case Tell Us?", *Perspectives of Psychological Science 12* (3): 374–381. DOI: 10.1177/1745691616679465.

Katz, D. 1967. "Editorial", *Journal of Personality and Social Psychology 7*: 341–344.

Katz, J. 1988. *Seductions of Crime: The Sensual and Moral Attractions of Doing Evil*. New York: Basic Books.

Kelley, H. H. 1992. "Common Sense Psychology and Scientific Psychology", *Annual Review of Psychology 43*: 1–23.

Kelley, H. H. 1999. "Fifty Years in Social Psychology: Some Reflections on the Individual-Group Problem". Pp. 35–46 in *Reflections on 100 Years of Experimental Social Psychology*, edited by A. Rodrigues and R. Levine. New York: Basic Books.

Kendrick, W. 1988. *The Secret Museum: Pornography in the Modern Culture*. New York: Viking Penguin.

Key, W. B. 1973. *Subliminal Seduction: Ad Media's Manipulation of Not so Innocent America*. Englewod Cliffs NJ: Prentice-Hall.

Klein, R. A. and 51 others. 2014. "Investigating Variation in Replicability: A 'Many Labs' Replication Project", *Social Psychology 45* (3): 142–152. DOI: 10.1027/1864-9335/a000178.

Kleinfeld, J. 1998. *The Myth that Schools Shortchange Girls: Social Science in the Service of Deception*. Washington, DC: Women's Freedom Network.

Kleinfeld, J. 1999. "Student Performance: Males Vs. Females", *Public Interest 134*: 3–20.

Klineberg, O. 1948. *Social Psychology*. New York: Holt.

Koch, S. 1992a. "Psychology Cannot Be a Coherent Science", *Psychology Today 14* (September):64, 66–68.

Koch, S. 1992b. "Wundt's Creature at Age Zero—and as Centenarian: Some Aspects of the Institutionalization of the 'New Psychology.'". Pp. 7–35 in *A Century of Psychology as Science*, edited by S. Koch and D. E. Leary. Washington, DC: American Psychology Association.

Krech, D. and R. S. Crutchfield 1948. *Theory and Problems of Social Psychology*. New York: McGraw-Hill.

Kuhn, T. 1970. *The Structure of Scientific Revolutions*. Chicago, IL: Chicago University Press.

Kutchinsky, B. 1991. "Pornography and Rape: Theory and Practice?", *International Journal of Law and Psychiatry 14* (*l*/*2*): 47–64.

LaCour, M. and D. Green 2014. "When Contact Changes Minds: An Experiment on Transmission of Support for Gay Equality," *Science 346* (6215): 1366–1369. DOI:10.1126/science.1256151.

LaPiere, R. T. and P. R. Farnsworth 1949. *Social Psychology* (3rd ed.). New York: McGraw-Hill.

Latané, B. and J. M. Darley 1968. "Group Inhibition of Bystander Intervention in Emergencies", *Journal of Personality and Social Psychology 10*: 215–221.

Latané, B. and J. M. Darley 1970. *The Unresponsive Bystander: Why Doesn't He Help?* New York: Appleton–Century–Crofts.

Latané, B. and S. Nida 1981. "Ten Years of Research on Group Size and Helping", *Psychological Bulletin 89*: 308–324.

Latané, B and J. Rodin 1969. "A Lady in Distress: Inhibiting Effects of Friends and Strangers on Bystander Intervention", *Journal of Experimental Social Psychology 5*: 189–202. DOI: 10.1016/0022-1031(69)90046-8.

Latour, B. and S. Woolgar 1979. *Laboratory Life*. London: Sage.

Learner, E. E. 1990. "Let's Take the Con Out of Econometrics". Pp. 29–49 in *Modeling Economic Series*, edited by C. W. J. Granger. Oxford: Clarendon.

Le Bon, G. 1895 *La psychologie des foules*. Translated and published in English in 1896 as *The Crowd: A study of the popular mind*, London: T.Fisher Unwin.

Le Texier, T. 2018. *Histoire D'un Mensonge: Enquête Sur L'expérience De Stanford*. Paris FR: Zones, Éditions La Découverte.

Le Texier, T. 2019. "Debunking the Stanford Prison Experiment", *American Psychologist 24* (7): 823–839.

Liebert, R. M. and J. Sprafkin 1988. *The Early Window: Effects of Television on Children and Youth* (3rd ed). New York: Pergamon.

Lemann, N. 2014. "A Call for Help: What the Kitty Genovese Story Really Means," *The New Yorker*, March 10th, online. Accessed 15 March 2014.

Levelt, W. J., M. E. Noort and P. Drenth 2012. *Flawed Science: The Fraudulent Reseqrch Practices of Social Psychologist Diederik Stapel*, Committee Report, University of Tilburg, online https://poolux.psychopool.tu-dresden.de/mdcfiles/gwp/Reale%20Fälle/Stapel%20-%20Final%20Report.pdf, accessed 21 February 2020.

Levy, L. 1974. "Awareness, Learning and the Beneficent Subject as Expert Witness". Pp. 187–197 in *The Experiment as a Social Occasion*, edited by P. Wuebben, B. Straits and G. Schulma. Berkeley, CA: Glendessary.

Levine, R.V. and A. Rodrigues (1999) "Afterword; Reflecting on Reflections". Pp. 215–230 in *Reflections on 100 Years of Experimental Social Psychology*, edited by A. Rodriques and R.V. Levine. New York: Basic Books.

Lewin, K. 1931. "The Conflict between Aristotelian and Galilean Modes of Thought in Contemporary Psychology", *Journal of General Psychology 5*: 141–176. online http://lchc.ucsd.edu/MCA/Mail/xmcamail.2012_07.dir/pdf1doBlYTMzb.pdf.

Lewin, K. 1951. *Field Theory in Social Science*. New York: Harper.

Lindesmith, A. R. and A. L. Strauss 1949. *Social Psychology*. New York: Dryden.

Linz, D. and E. Donnerstein 1988. "The Methods and Merits of Pornography Research", *Journal of Communication 38* (2): 180–184.

Linz, D., S. Penrod and E. Donnerstein 1987. "The Attorney General's Commission on Pornography: The Gap between 'Findings' and Fact", *American Bar Foundation Research Journal* Fall: 713–736.

Loftus, E. 1993. "The Reality of Repressed Memories", *American Psychologist 48* (5): 518–537.

Loftus, E. and K. Ketcham 1994. *The Myth of Repressed Memory: False Memories and Allegations of Sexual Abuse*. New York: St. Martin's.

Lowy, S. 1944. *Man and His Fellowmen: Modern Chapters on Social Psychology.* London: Kegan Paul, Trench, Trubner & Co.

Luker, K. 1984. *Abortion and the Politics of Motherhood*. Berkeley and Los Angeles, CA: University of California Press.

Luria, Z. 1986. “A Methodological Critique”, *Signs: Journal of Women in Culture and Society 11*: 316–321.

Maccoby, E. 1985. “Social Groupings in Childhood: Their Relationship to Prosocial and Antisocial Behavior in Boys and Girls”. Pp. 263–285 in *Development of Antisocial Behavior and Prosocial Behavior: Theories, Research and Issues*, edited by D. Olweus, J. Block and M. Radke-Yarrow. San Diego, CA: Academic Press.

MacKinnon, C. A. 1985. “Pornography, Civil Rights and Speech”, *Harvard Civil Rights-Civil Liberties Law Review 20*: 2–70.

Maeder, J. 2017. “How the Murder of Kitty Genovese Rattled the Conscience of New York City” *The New York Daily News*, August 14, 2017. Accessed on 19 December 2019.

Malamuth, N. M. and J. V. P. Check 1981. “The Effects of Mass Media Exposure on Acceptance of Violence against Women: A Field Experiment”, *Journal of Research in Personality 15*: 436–446.

Mannheim, K. 1954. *Ideology and Utopia*. Translated by Louis Wirth and Edward Shils. London: Routledge and Kegan Paul.

Manning, R., M. Levine and A. Collins 2007. “The Kitty Genovese Murder and the Social Psychology of Helping”, *American Psychologist 62* (6): 555–562.

Mantel, D. M. 1971. “The Potential for Violence in Germany”, *Journal of Social Issues 27* (4): 110–111.

Mayo, E. 1933. *The Human Problems of an Industrial Civilisation*. New York: Macmillan.

Mayo, E. 1939. “Preface”. Pp. xi–xiv in *Management and the Worker*, edited by F. J. Roeth-lisberger and W. J. Dickson. Cambridge, MA: Harvard University Press.

McCambridge, J., J. Witton and D.R. Elbourne 2014. “Systematic review of the Hawthorne Effect: New Concepts are needed to study research participation effects,” *Journal of Clinical Epistemology* 67 (3): 267–77.

McDougall, W. 1919. *An Introduction to Social Psychology* (14th rev ed.). London: Methuen.

Mead, G. H. 1934. *Mind, Self and Society from the Standpoint of a Social Behaviorist*. Chicago, IL: University of Chicago Press.

Mednick, M. T. 1989. “On the Politics of Psychological Constructs: Stop the Bandwagon, I Want to Get Off”, *American Psychologist 44* (8): 1118–1123.

Merton, R. K. 1948. “The Self-Fulfilling Prophecy”, *Antioch Review 8*: 193–210.

Milgram, S. 1963. “Behavioral Study of Obedience”, *Journal of Abnormal and Social Psychology 67* (4): 371–378.

Milgram, S. 1965. “Liberating Effects of Group Pressure”, *Journal of Personality and Social Psychology 19*: 137–143.

Milgram, S. 1974. *Obedience to Authority*. New York: Harper and Row.

Miller, A. G. 1986. *The Obedience Experiments: A Case Study of Controversy in Social Science*. New York: Praeger.

Miller, G. A. 1992. “The Consitutive Problem of Social Psychology”. Pp. 40–46 in *A Century of Psychology as Science*, edited by S. Koch and D. E. Leary. Washington, DC: American Psychology Association.

Mischel, W. 2014. *The Marshmallow Test: Mastering Self-control*. New York: Little, Brown and Co.

Mixon, D. 1971. “Beyond Deception”, *Journal for the Theory of Social Behaviour 2* (2): 145–177.

Moffitt, T. E. 1993. “Adolescent-Limited and Life-Course Persistent Antisocial Behavior: A Developmental Taxonomy”, *Psychological Review 100*: 674–701.

Molden, D. C. 2014. “Understanding Priming Effects in Social Psychology: What Is ‘Social Priming’ and How Does It Work?”, *Social Cognition 32*: 1–14.

Moscovici, S. 1972. “Society and Theory in Social Psychology”. Pp. 17–96 in *The Context of Social Psychology: A Critical Assessment*, edited by J. Israel and H. Tajfel. New York: Academic Press.

Mulvey, E. P. and J. L. Haugaard 1986. *Pornography and Public Health*. Washington, DC: U.S. Department of Health and Human Services.

Murchison, C. A. 1935. *A Handbook of Social Psychology*. New York: Russell & Russell.

Murphy, G., L. B. Murphy and T. M. Newcomb 1937. *Experimental Social Psychology: An Interpretation of Research upon the Socialization of the Individual* (rev. ed.). New York: Harper.

Myer, M. N. and C. Chabris 2014. “Why Psychologists’food Fight Matters: ‘Important Findings’ Haven’t Been Replicated”, Slate, July 31 online https://slate.com/technology/2014/07/replication-controversy-in-psychology-bullying-file-drawer-effect-blog-posts-repligate.html, accessed 15 January 2020.

National Council for Research on Women. 1998. *The Girls Report: What We Know and Need to Know about Growing up Female*. Washington, DC: NCRW.

National Institute of Mental Health. 1982. *Television and Behaviour: Ten Years of Scientific Progress and Implications for the Eighties*. Washington, DC: NIMH.

NCES. 2019. *Digest of Educational Statistics 2018*, (54thed.), Washington DC: National Center for Educational Statistics, U.S. Department of Education. NCES 2020-009, https://nces.ed.gov/pubsearch/pubsinfo.asp?pubid=2020009, accessed on 13 February 2020.

Newcomb, T. M. and E. L. Hartley (Eds.). 1947. *Readings in Social Psychology*. New York: Holt.

Nicholson, I. 2011. “Torture at Yale”, *Experimental Subjects, Laboratory Torment and the ‘Rehabilitation’ of Milgram’s ‘Obedience to Authority’, Theory and Psychology 21* (6): 737–761. DOI: 10.1177/0959354311420199.

Nicholson, I. 2015. “The Normalization of Torment: Producing and Managing Anguish in Milgram’s ‘Obedience’ Laboratory”, *Theory and Psychology 25* (5): 639–656. DOI: 10.1177/0959354315605393.

Nosek, B. A. and D. Lakens 2014. “Registered Reports: A Method to Increase the Credibility of Published Results”, *Social Psychology 45* (3): 137–141. DOI: 10.1027/1864-9335/a000192.

Olweus, D. 1979. “’Stability of Aggressive Reaction Patterns in Males’: A Review”, *Psychological Bulletin 86*: 852–875.

Ongley, S. F., M. Nola and T. Maiti 2014. “Children’s Giving: Moral Reasoning and Moral Emotions in the Development of Donation Behaviors”, *Frontiers in Psychology 5*: 1–8. DOI: 10.3389/fpsyg.2014.00458.

Open Science Collaboration. 2012. “An Open, Large-scale, Collaborative Effort to Estimate the Reproducibility of Psychological Science”, *Perspectives on Psychological Science 7*: 657–660. DOI: 10.1177/1745691612462588.

Open Science Collaboration. 2015. “Estimating the Reproducibility of Psychological Science”, *Science 349* (6251) 28 August online https://science.sciencemag.org/content/349/6251/aac4716.full?ijkey=1xgFoCnpLswpk&keytype=ref&siteid=sci, 10.1126/science.aac4716.

Orne, M. T. 1962. “On the Social Psychology of the Psychological Experiment with Particular Reference to Demand Characteristics and Their Implications”, *American Psychologist 17*: 777–783.

Orne, M. T. and C. H. Holland 1968. “On the Ecological Validity of Laboratory Deceptions”, *International Journal of Psychiatry 6*: 282–293.

Orne, M.T. and K.E. Schneibe (1964). “The contribution of non-deprivation factors in the production of sensory deprivation effects,” *Journal of Abnormal and Social Psychology* 68: 3–13.

Parsons, H. M. 1974. “What Happened at Hawthorne?”, *Science 183*: 922–932.

Pashler, H. and E.-J. Wagenmakers 2012. “Editors’ Introduction to the Special Edition on Replicability in Psychological Science: A Crisis of Confidence?”, *Perspectives on Psychological Science 7* (6): 528–530. DOI: 10.1177/1745691612465253.

Patten, S. 1977a. “The Case That Milgram Makes”, *Philosophical Review 86* (3): 350–364.

Patten, S. 1977b. “Milgram’s Shocking Experiments”, *Philosophy 52* (4): 425–450.

Pepitone, A. 1999. “Historical Sketches and Critical Commentary about Social Psychology in the Golden Age”. Pp. 170–199 in *Reflections on 100 Years of Experimental Social Psychology*, edited by A. Rodrigues and R. Levine. New York: Basic Books.

Perry, G. 2012. *Behind the Shock Machine: The Untold Story of the Notorious Milgram Psychology Experiments*. Melbourne and London: Scribe Books, Revised: New York: The New Press, 2013.

Perry, G. 2018. *The Lost Boys: Inside Muzafer Sherif’s Robbers Cave Experiment*. Melbourne and London: Scribe Books.

Perry, G., A. Brannigan, R. Wanner and H. Stam 2020. “Credibility and Incredulity in Milgram’s Obedience Experiments: A Reanalysis of an Unpublished Test”, *Social Psychology Quarterly*, DOI: 10.1177/0190272519861952.

Perry, W. J. 1935. *The Primordial Ocean: An Introductory Contribution to Social Psychology*. London: Methuen.

Petersen, J. L. and J. S. Hyde 2010. “A Meta-analytic Review of Research on Gender Differences in Sexuality: 1993–2007”, *Psychological Bulletin 136* (1): 21–38.

Pevalin, D. J., T. J. Wade and A. Brannigan 2003. “Precursors, Consequences and Implications for Stability and Change in Pre-adolescent Antisocial Behaviors”, *Prevention Science 4* (2): 123–136.

Pfungst, O. [1911] 1965. *Clever Hans (The Horse of Mr. Von Osten): A Contribution to Experimental, Animal, and Human Psychology*. Translated by C. L. Rahn. New York: Holt, Rinehart and Winston.

Philpot, R., L. S. Liebst, M. Levine, W. Bernasco and M. R. Lindegaard 2019. “Would I Be Helped? Cross-national CCTV Footage Shows that Intervention Is the Norm in Public Conflicts”, *American Psychologist*. Advance online publication. DOI: https://doi.org/10.1037/amp0000469. Accessed on 21 December 2019.

Popper, K. 1960. *The Poverty of Historicism* (2nded.). Boston MA: Beacon Press.

Popper, K. R. 1959. *The Logic of Scientific Discovery*. New York: Harper Torch-backs.

Popper, K. R. [1961] 1976. “The Logic of the Social Sciences”. Pp. 87–104 in *The Positivist Dispute in German Sociology*, edited by T. W. Adorno, H. Albert, R. Dahrendorf, J. Habermas, H. Pilot and K. L. Popper. Translated by Glyn Adey and David Frisby. London: Heinemann.

Porn Hub. 2019. www.pornhub.com/insights/2019-year-in-review. Accessed 28 January 2020.

Porter, C. 2012. “The Hawthorne Effect Today,” *Industrial Management 54* (3) May/June. Online. www.questia.com/magazine/1P3-2824722571/the-hawthorne-effect-today. Accessed 2 January 2020.

Posner, R. A. 1988. *Law and Literature: A Misunderstood Relation*. Cambridge, MA: Harvard University Press.

Prentice, D. A. and D. T. Miller 2006. “Essentializing Differences between Women and Men”, *Psychological Science 17* (2): 129–135.

Proshansky, H. and B. Seidenberg (Eds.). 1965. *Basic Studies in Social Psychology.* New York: Holt, Rinehart and Winston.

Raudenbush, S. W. 1984. “Magnitude of Teacher Expectancy Effects on Pupil IQ as A Function of the Credibility of Expectancy Induction: A Synthesis of Findings from 18 Experiments”, *Journal of Educational Psychology 76*: 85–97.

Raven, B. 1999. “Reflections on Interpersonal Influence and Social Power in Experimental Social Psychology”. Pp. 114–134 in *Reflections on 100 Years of Experimental Social Psychology*, edited by A. Rodrigues and R. Levine. New York: Basic Books.

Ridley, M. 2003. *Nature via Nurture: Genes, Experience and What Makes Us Human*. Toronto, CA: Harper Collins.

Ring, K. 1967. “Experimental Social Psychology: Some Sober Questions About Some Frivolous Values”, *Journal of Experimental Social Psychology 3*: 113–123.

Robinson, V. P. 1930. *A Changing Psychology in Social Case Work*. Chapel Hill, NC: University of North Carolina Press.

Rodrigues, A. and R. V. Levine (Eds.). 1999. *Reflections on 100 Years of Experimental Social Psychology.* New York: Basic Books.

Roethlisberger, F. J. and W. J. Dickson 1939. *Management and the Worker.* Preface by Elton Mayo. Cambridge MA: Harvard University Press.

Rose, H. and S. Rose (Eds.). 2000. *Alas, Poor Darwin: Arguments Against Evolutionary Psychology.* London: Jonathan Cape.

Rosenhan, D. L. 1973. “Being Sane in Insane Places”, *Science 179*: 250–258.

Rosenthal, A. M. 1964. *Thirty-eight Witnesses*. Berkeley: University of California Press (Revised 1999).

Rosenthal, R., S. S. Baratz and C. M. Hall 1974. “Teacher Behaviour, Teacher Expectations, and Gains in Pupils’ Rated Creativity”, *Journal of Genetic Psychology 124*: 115–121.

Rosenthal, R. 1966. *Experimenter Effects in Behavioral Research*. New York: Meredith.

Rosenthal, R. 1969. “Empirical Vs. Decreed Validation of Clocks and Tests”, *American Educational Research Journal 6*: 689–691.

Rosenthal, R. 1985. “From Unconscious Experimenter Bias to Teacher Expectancy Effects”. Pp. 37–65 in *Teacher Expectancies*, edited by J. B. Dusek. Hillsdale, NJ: Lawrence Erlbaum Associates.

Rosenthal, R. 1987. *Judgment Studies: Design, Analysis, and Meta-Analysis*. New York: Cambridge University Press.

Rosenthal, R. and L. Jacobson 1968. *Pygmalion in the Classroom: Teacher Expectations and Pupils’Intellectual Development*. New York: Holt, Rinehart and Winston.

Rosenthal, R. and L. Jacobson 1969. “Teacher Expectations for the Disadvantaged”, *Scientific American 218* (4): 19–23.

Rosenwald, G. 1986. “Why Operationism Doesn’t Go Away: The Extrascientific Incentives of Social-psychological Research”, *Philosophy of the Social Sciences 16* (3): 303–330.

Ross, E. A. 1908. *Social Psychology: An Outline and Source Book*. New York: Macmillan.

Russell, D. 1993. “The Experts Cop Out”. Pp. 151–167 in *Making Violence Sexy*, edited by D. Russell. Williston, VT: Teachers College Press.

Russell, N. 2018. *Understanding Willing Participants: Milgram’s Obedience Experiments and the Holocaust*. Two Volumes. London: Palgrave.

Russell, N. J. C. 2011. "Milgram's Obedience to Authority Experiments: Origins and Early Evolution", *British Journal of Social Psychology 50* (1): 140–162. DOI: 10.1348/014466610X492205.

Sabini, J. 1986. "Stanley Milgram (1933–1984)", (obituary). *American Psychologist 41* (12): 1378–1379.

Satcher, D. 1999. *Youth Violence: A Report of the Surgeon General*. Washington, DC: Public Health Service.

Savage, J. and C. Yancey 2008. "The Effects of Media Violence on Criminal Aggression: A Meta-analysis", *Criminal Justice and Behavior 35* (6): 772–791. DOI: 10.1177/0093854808316487.

Schachter, S. 1959. *The Psychology of Affiliation*. Stanford: Stanford University Press.

Schachter, S. 1971. *Emotion, Obesity and Crime*. New York: Academic Press.

Schachter, S. 1980. "Non-Psychological Explanations of Behavior". Pp. 131–157 in *Retrospections on Social Psychology*, edited by L. Festinger. New York: Oxford.

Schachter, S. and J. E. Singer 1962. "Cognitive, Social and Physiological Determinants of Emotional State", *Psychological Review 69*: 379–399.

Schipani, V. 2018. "The Truth about Media Violence," *Yancey report a meta-analysis*, March 15, Online https://undark.org/2018/03/15/the-truth-about-media-violence. Accessed 31 January 2020.

Schimmack, U., M. Heene and K. Kesevan 2017. "Reconstruction of a Train Wreck: How Priming Research Went off the Rails", Replicability-Index online blog, https://replicationindex.com/2017/02/02/reconstruction-of-a-train-wreck-how-priming-research-went-of-the-rails/. Accessed 3 February 2020.

Schmitt, D. P. 2019. "Are Men More Helpful, Altruistic or Chivalrous than Women?" *Psychology Today* blog, posted 10 March 2016, updated 6 September 2019. Online www.psychologytoday.com/us/blog/sexual-personalities/201603/are-men-more-helpful-altruistic-or-chivalrous-women, accessed 14 February 2020.

Schnall, S., J. Benton and S. Harvey. 2008. "With a Clean Conscience: Cleanliness Reduces the Severity of Moral Judgments", *Psychological Science 19* 1219–1222. DOI: 10.1111/j.1467-9280.2008.02227.x

Schumacher, H. 2019. "Why More Men than Women Die by Suicide," London: BBC Future online, www.bbc.com/future/article/20190313-why-more-men-kill-themselves-than-women, accessed 13 February 2020.

Scott, J. and L. A. Schwalm 1988. "Rape Rates and the Circulation Rates of Adult Magazines", *Journal of Sex Research 24*: 241–250.

Sedivy, J. 2011. "Subliminal Seduction Gets a Second Glance" Psychology Today, January 27 online, www.psychologytoday.com/us/blog/sold-language/201101/subliminal-seduction-gets-second-glance, accessed 4 March 2020.

Shanks, D. R., B. R. Newell, E. H. Lee, D. Balakrishnan, L. Ekelund, Z. Cenac, F. Kavvadia and C. Moore 2013. "Priming Intelligent Behavior: An Elusive Phenomenon", *PLoS ONE 8* (4): e56515. DOI: 10.1371/journal.pone.0056515.

Sherif, M. 1935. "A Study of Some Social Factors in Perception", *Archives of Psychology 27* (187): 1–60.

Sherif, M. [1936] 1965. *The Psychology of Social Norms*. Introduction by Gardner Murphy. New York: Octagon.

Sherif, M. 1937. "An Experimental Approach to the Study of Attitudes", *Sociometry 1*: 90–98.

Sherif, M. 1956. "Experiments in Group Conflict", *Scientific American 195* (5): 54–59.

Sherif, M. B., J. White and O. J. Harvey 1955. “Status in Experimentally Produced Groups”, *American Journal of Sociology 60* (4): 370–379.

Silverman, I. 1971. “Crisis in Social Psychology: The Relevance of Relevance”, *American Psychologist 26*: 583–604.

Silverman, I. 1977. “Why Social Psychology Fails”, *Canadian Psychological Review 18* (4): 353–408.

Simmons, J. P., L. D. Nelson and U. Simonsohn 2011. “False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis”, *Psychological Science 22* (11): 1359–1366. DOI: 10.1177/0956797611417632.

Simonsohn, U. 2013. “Just Post It: The Lesson from Two Cases of Fabricated Data Detected by Statistics Alone”, *Psychological Science 24* (10): 1875–1888. DOI: 10.1177/0956797613480366.

Skinner, B. F. [1948] 1976. *Walden Two*. New York: Macmillan.

Skocai, M., M. Filipic, J. Petkovic and S. Novak 2011. “Titanium Dioxide in Our Everyday Life; Is It Safe?” *Radiology and Oncology 45* (4): 227–247. DOI: 10.2478/v10019-011-0037-0.

Slocombe, C. S. 1940. “Million Dollar Research”, *Personnel Journal 18*: 162–172.

Smith, M. B. 1972. “Is Experimental Social Psychology Advancing?”, *Journal of Experimental Social Psychology 8*: 86–96.

Smith, M. B. 1973. “Is Psychology Relevant to New Priorities?”, *American Psychologist 28*: 463–471.

Smith, M. L. 1980. “Teacher Expectations”, *Evaluation in Education: An International Journal 4*: 53–55.

Smoke, K. 1935. “The Present Status of Social Psychology in America”, *Psychological Review 42*: 537–553.

Sobol, M. G. 1959. “Panel Mortality and Panel Bias”, *Journal of the American Statistical Association 54* (285): 52–68.

Sokal, A. 1996a. “A Physicist Experiments with Cultural Studies”, *Lingua Franca 6* (4): 62–64.

Sokal, A. 1996b. “Transgressing the Boundaries: Towards a Transformative Hermeneutics of Quantum Gravity”, *Social Text 14* (1, 2): 217–252.

Solomon, J. 2016. *The Witness*, www.pbs.org/independentlens/films/witness/. DVD 90 minutes. www.rottentomatoes.com/m/the_witness_2016.

Sommers, C. H. 1994. *Who Stole Feminism? How Women Have Betrayed Women*. New York: Simon & Schuster.

Sommers, C. H. 2000a. *The War Against Boys*. New York: Simon and Schuster.

Sommers, C. H. 2000b. “The War Against Boys: How Misguided Feminism Is Harming Our Young Men”, *Atlantic Monthly 285* (5): 59–74.

Sorokin, P. 1954. *Fads and Foibles in Modern Sociology*. Chicago: Regnery.

SSRP. 2016. Social Science Replication Project Homepage. www.socialsciencesreplicationproject.com. Accessed 15 February 2020.

Stadler, D. R. 2008. “Revisiting the Issue of Safety in Numbers: The Likelihood of Receiving Help from a Group”, *Social Influence 3* (1): 24–33.

Stadler, D. R. 2019. “New Study Suggests that Bystander Apathy Is Not the Norm,” *Psychology Today online*, www.psychologytoday.com/ca/blog/bias-fundamentals/201907/new-study-suggests-bystander-apathy-is-not-the-norm. Accessed 20 December 2019.

Stam, H. J., H. L. Radtke and I. Lubek 1998. “Repopulating Social Psychology Texts”. Pp. 153–186 in *Reconstructing the Psychological Subject*, edited by B. M. Bayer and J. Shotter. London: Sage.

Stam, H. J., H. L. Radtke and I. Lubek 2000. "Strains in Experimental Social Psychology: A Textual Analysis of the Development of Experimentation in Social Psychology", *Journal of the History of the Behavioral Sciences 36* (4): 365–383.

Stapel, D. 2012. *Ontsporing*, Translated (in part) by Nicholas J.L. Brown in 2014 as *Faking Science: A True Story of Academic Frau*d, Online http://nick.brown.free.fr/stapel/FakingScience-20161115.pdf

StatsCan. 2012. Suicide Rates: An Overview (Tanya Navaneelan), Catalogue no. 82-624-X, Ottawa: Statistics Canada, www150.statcan.gc.ca/n1/en/pub/82-624-x/2012001/article/11696-eng.pdf?st=duIxOImc, accessed 13 February 2020.

StatsCan. 2016. *Women and Education: Qualifications, Skills and Technology* (Sarah J. Ferguson), Catalogue no. 89-503-X, Ottawa: Statistics Canada, www150.statcan.gc.ca/n1/en/pub/89-503-x/2015001/article/14640-eng.pdf?st=hAh7sPS8, accessed 13 February 2020.

Sternberg, S. 2020. "Are Those Breakthrough Heart Studies Really to Be Trusted?" US News March 2 online www.usnews.com/news/health-news/articles/2020-03-02/are-those-breakthrough-heart-studies-really-to-be-trusted, accessed 3 2 March 2020.

Strieker, L. 1967. "The True Deceiver", *Psychological Bulletin 68* (1): 13–20.

Surgeon General's Scientific Advisory Committee. 1972. *Television and Growing Up: The Impact of Televised Violence*. Washington, DC: USPGO.

Świątkowski, W. and B. Dompnier 2017. "Replicability Crisi in Social Psychology: Looking at the past to Find Pathways for the Future", *International Review of Social Psychology 30* (1): 111–124. DOI: https://doi.org/10.5334/irsp.66.

Tannenbaum, P. H. 1972. "Studies in Film- and Television-Mediated Arousal and Aggression: A Progress Report". Pp. 309–350 in *Television and Social Behavior: Vol. 5, Television's Effects: Further Explorations*, edited by George A. Comstock, Eli A. Rubinstein and John P. Murray. Washington, DC: USGPO.

Tavris, C. 2002. "The High Cost of Skepticism." *Skeptical Inquirer*, July-August (on-line).

Thoma, S. 1986. "Estimating Gender Differences in the Comprehension and Preferences of Moral Issues", *Developmental Review 6*: 165–180.

Thorndike, R. L. 1968. "Review of Pygmalion in the Classroom", *American Educational Research Journal 5* (4): 708–711. Reprinted in Janet D. Elashoff and Richard E. Snow (eds.) *Pygmalion Reconsidered*, Worthington. Ohio: Charles Jones, 1971.

Timmermans, A. C., C. M. Rubie-Davies and C. Rjosk 2018. "Pygmalion's 50th Anniversary: The State of the Art in Teacher Expectation Research", *Educational Research and Evaluation 24* (3–5): 91–98.

Tolkien, J. R. R. 1966. *The Lord of the Rings*. London: Grafton, Pp. 65–68.

Tooby, J. and L. Cosmides 1996. "The Psychological Foundations of Culture". Pp. 19–131 in *The Adapted Mind: Evolutionary Psychology and the Generation of Culture*, edited by J. Barkow, L. Cosmides and J. Tooby. New York: Oxford University Press.

Triplett, N. 1897. "The Dynamogenic Factors in Pace-Making and Competition", *American Journal of Psychology 9*: 507–532.

Twain, M. 1885. *The Adventures of Huckleberry Finn*. New York: Random House.

van Kolfschooten, F. 2015. "Report Further Incriminates Social Psychologist Jens Förster" *Science* ScienceMag.org 2 June 2015online www.sciencemag.org/news/2015/06/report-further-incriminates-social-psychologist-jens-f-rster, accessed 15 February 2020.

Vonks, J. 2012. "Fraud Case Dirk Smeesters" online blog www.jennifervonk.com/uploads/7/7/3/2/7732985/smeesterscase.pdf, accessed 20 February 2020.

Vyse, S. 2017. "P-Hackerr Confessions: Daryl Bem and Me" *Skeptical Inquirer* online https://stuartvyse.com/2017/06/14/p-hacker-confessions-daryl-bem-me/, accessed on 1 February 2020.

Walker, L. 1984. “Sex Differences in the Development of Moral Reasoning: A Critical Review”, *Child Development 55*: 677–691.

Walker, L. J. 2006. “Gender and Morality”. Pp. 93–115 in *Handbook of Moral Psychology*, edited by M. Killen and J. Smetana. Mahwah NJ: Erlbaum.

Weber, M. 1958. *The Protestant Ethic and the Spirit of Capitalism*. 1920-21 edition translated from the German by Talcott Parsons, Introduction by Anthony Giddens. New York: Scribner.

Webster's Seventh New Collegiate Dictionary. 1977. Toronto, CA: Thomas Allen and Sons Limited.

Wertham, F. 1954. *Seduction of the Innocent*. New York: Rinehart.

Wheeler, L. 1987. “Social Comparison, Behavioral Contagion, and the Naturalistic Study of Social Interaction”. Pp. 46–65 in *A Distinctive Approach to Psychological Research: The Influence of Stanley Schachter*, edited by N. E. Grunberg, R. E. Nisbett, J. Rodin and J. E. Singer. Hillsdale, NJ: Lawrence Erlbaum Associates.

Whitehead, T. N. 1938. *The Industrial Worker*. Cambridge, MA: Harvard University Press.

Wiegman, O., M. Kuttschreuter and B. Baarda 1992. “A Longitudinal Study of the Effects of Television Viewing on Aggression and Prosocial Behaviors”, *British Journal of Social Psychology 31*: 147–164.

Wilson, J. Q. and R. J. Herrnstein 1985. *Crime and Human Nature: The Definitive Study of the Causes of Crime*. New York: Simon and Schuster.

Wineburg, S. S. 1987a. “Does Research Count in the Lives of Behavioral Scientists?”, *Educational Researcher 16* (9): 42–44.

Wineburg, S. S. 1987b. “The Self-Fulfillment of a Self-Fulfilling Prophecy”, *Educational Researcher 16* (9): 28–37.

Wittgenstein, L. 1922. *Tractatus Logico-Philosophicus*. London: Routledge and Kegan Paul.

Wittgenstein, L. 1951. *Philosophical Investigations*. Oxford: Macmillan.

Wuebben, P., B. Straits and G. Schulman (Eds.). 1974. *The Experiment as a Social Occasion*. Berkeley, CA: Glendessary.

Wurtzel, A. and G. Lometti. 1984. “Smoking Out the Critics.” *Society* (September/October): 36–40.

Yglesia, M. 2018. “The Bell Curve Is about Policy. And It's Wrong,” *Vox* 10 April 2018. www.vox.com/2018/4/10/17182692/bell-curve-charles-murray-policy-wrong. Retrieved 6 January 2020.

Yong, E. 2018. “Psychology's Replication Crisis Is Running Out of Excuses”, *The Atlantic* online, www.theatlantic.com/science/archive/2018/11/psychologys-replication-crisis-real/576223/. Accessed 12 January 2020.

You, D., Y. Maeda and M. Bedeau 2011. “Gender Differences in Moral Sensitivity: A Meta-analysis”, *Ethics and Behavior 21* (4): 263–282.

Zajonc, R. B. 1997. “One Hundred Years of Rationality Assumptions in Social Psychology”. Pp. 200–214 in *Reflections on 100 Years of Experimental Social Psychology*, edited by Aroldo Rodrigues and Robert V. Levine. New York: Basic Books.

Zhong, C. B., B. Strejcek and N. Sivanathan 2010. “A Clean Self Can Render Harsh Moral Judgment”, *Journal of Experimental Social Psychology 46*: 859–862. DOI: 10.1016/j.jesp.2010.04.003.

Zillmann, D. and W. J. Bryant 1982. “Pornography, Sexual Callousness and the Trivialization of Rape”, *Journal of Communication 34*: 10–21.

Zillmann, D. and W. J. Bryant 1984. “Effects of Massive Exposure to Pornography”. Pp. 115–138 in *Pornography and Sexual Aggression*, edited by N. Malamuth and E. Donnerstein. Orlando, FL: Academic Press.

Zillmann, D. 1991. “Television Viewing and Physiological Arousal”. Pp. 103–133 in *Responding to the Screen: Receptions and Reaction Processes*, edited by J. Bryant and D. Zillmann. Hillsdale, NJ: Lawrence Erlbaum Associates.

Zimbardo, P. 1972. “Comment: Pathology of Imprisonment”, *Society* April 9 (6): 4–8.

Zimbardo, P. 1999. “Experimental Social Psychology: Behaviorism with Minds and Matters”. Pp. 135–157 in *Reflections on 100 Years of Experimental Social Psychology*, edited by A. Rodrigues and R. Levine. New York: Basic Books.

Zimbardo, P. 2007. *The Lucifer Effect: Understanding How Good People Turn Evil*. New York: Random House.

Zimbardo, P. 2018. “Statement from Philip Zimbardo: Response to Recent Criticisms of the Stanford Prison Experiment” online www.prisonexp.org/response, retrieved 30 April 2020.

Znaniecki, F. 1925. *The Laws of Social Psychology*. New York: Russell & Russell.

Index