SOCIUS

Original Article

AMERICAN SOCIOLOGICAL ASSOCIATION

Socius: Sociological Research for a Dynamic World
Volume 2: 1–10
© The Author(s) 2016
DOI: 10.1177/2378023115625071
srd.sagepub.com
SAGE

The Baby Factory: Difficult Research Objects, Disciplinary Standards, and the Production of Statistical Significance

David Peterson[1]

Abstract

Science studies scholars have shown that the management of natural complexity in lab settings is accomplished through a mixture of technological standardization and tacit knowledge by lab workers. Yet these strategies are not available to researchers who study difficult research objects. Using 16 months of ethnographic data from three laboratories that conduct experiments on infants and toddlers, the author shows how psychologists produce statistically significant results under challenging circumstances by using strategies that enable them to bridge the distance between an uncontrollable research object and a professional culture that prizes methodological rigor. This research raises important questions regarding the value of restrictive evidential cultures in challenging research environments.

Keywords

laboratory ethnography, science and knowledge, standardization, sociology of psychology

> Never work with children or animals.
>
> —W. C. Fields

It is a truism in science studies that scientific labs are interpretive and messy compared with the clean argumentation in scientific papers (Gilbert and Mulkay 1984; Holton 1978; Knorr Cetina 1995; Woolgar 1982). Early lab ethnographies argued that scientific research was produced through social negotiations within circumstances of uncontrollable contingency rather than unambiguous reflections of objective nature (Knorr Cetina 1983; Lynch 1985). However, like all truisms, the belief that labs are messy and contingent places can become a simplistic generalization.

Although all experimental laboratories may wrangle with environmental variability, the specifics and extent of this challenge differ in ways that scholars have yet to acknowledge. If, as Knorr Cetina (1995:145) argued, laboratories are defined by their ability to transplant, transform, and manipulate research objects, then it follows that any field that faces systematic constraints on any or all of these abilities would face unique difficulties in laboratory work. These, in turn, would influence the development of the entire field.

Outlining these constraints requires the researcher to investigate how a challenging research object or context produces *typical* situations in which researchers are forced to compensate or compromise. This goal requires new forms of research methods. Because any individual lab may be unethical or idiosyncratic, the ethnographer must observe multiple labs to understand which research challenges and solutions are common in the field.

In this article, I conduct three ethnographies of a particularly challenging research environment, developmental psychology labs that study infants and toddlers, to illustrate how researchers navigate a path between a difficult research object and a demanding disciplinary culture. Ultimately, I argue that developmental psychologists meet disciplinary requirements through a set of strategies that bend results toward statistical significance. Because these strategies also increase the risk of false positives, I argue that developmental psychologists counteract a problematic literature through the development of a local culture of evaluation that contextualizes findings within multiple streams of evidence.

[1]Northwestern University, Evanston, IL, USA

Corresponding Author:
David Peterson, Northwestern University, 1810 Chicago Avenue, Evanston, IL 60208, USA
Email: davidpeterson@u.northwestern.edu

Taming Natural Variability

The struggle against natural variability is one of the foundations of modern science. Science studies researchers have outlined several ways scientists attempt to produce stable and predictable lab conditions.

Standardization is one of the best understood strategies (e.g., Fujimura 1992; Jordan and Lynch 1998; Timmermans and Epstein 2010). This occurs along several dimensions. Latour (1983) showed how the successful extension of scientific research outside of the laboratory involved exporting the material culture of the lab—the machines and processes that standardize research objects—into new environments. Other research illustrated how important standardization was for the objects of research themselves. Daston and Galison (1992:85) argued that sciences are organized around "working objects," which are opposed to the highly variable objects of nature. In this vein, Kohler (1999) showed how the standardization of the fruit fly was foundational for the development of genetic research, and Epstein (2007) illustrated how early medical researchers attempted to accomplish object standardization by limiting admittance into test trials to a uniform population (i.e., middle-aged, white men).

However, taming natural variation through the standardization of research environments and objects is often not enough to ensure successful experimentation. A second strand of literature has focused on the intensive training that researchers must undergo to get experiments to work. Learning how to produce successful experiments (Collins 1974, 1985; Delamont and Atkinson 2001) is not something that can be explicitly codified and separated from actual practice. Instead, scholars have suggested that successful scientific practice requires some embodied "skill" or tacit "knowledge" that can be learned only through hands-on practice and local interaction (Collins 2001; 2010; MacKenzie and Spinardi 1995).

Thus, natural variability is made manageable through preexperimental standardization, or it is managed in situ by skilled experimenters. Through these methods, natural objects are transformed into something more predictable and orderly and, thus, something capable of linking diverse research sites and establishing a productive science. However, these strategies depend on a particular class of research objects: those that can be manipulated or mastered. This is not always possible. When the object of study is truly an "object"—a mere thing—there are few restrictions on the manipulations that can be used to make it suitable for lab science. Even animals can become mere objects. However, modern research involving humans is more restrictive, because ethical and legal concerns limit the possibilities for manipulation.

Untamable Variability in Developmental Psychology

Psychology in the early twentieth century was largely the study of the standardized white rat (Lemov 2005). It was assumed that they represented a simplified model of human behavior while offering "experimental convenience and control" that human studies could not match (Skinner 1938). They were inexpensive and abundant, and they could be freely manipulated. However, as the limitations of equating mouse and human behavior became increasingly apparent, psychologists began to experiment more on human subjects (Greenwood 1999).

Psychological research on humans has typically sought to tame their variability through two methods. First, psychologists present subjects with highly controlled testing environments. For instance, in Milgram's (1963) famous experiment, researchers standardized every aspect of the environment: the room, the electroshock generator, and scripts for both experimenter and confederate. However, because human behavior even in standardized environments remains highly variable, psychologists further enhanced predictability by providing detailed instructions to limit possible reactions and produce interpretable responses. The subjects were expected to standardize themselves.

Psychological research on human infants represents a particularly challenging limit case in the struggle over natural variability. Psychologists who work with nonhuman animals can breed them, control all aspects of their environment, restrain them, and run an unlimited number of studies on them. Adult human subjects, on the other hand, can usually be convinced to participate with the experimenter's wishes, eliminating the need for physical control. Infant subjects occupy a liminal space because they are as unpredictable and resistant to instruction as animals and yet bear the inviolable rights of human beings. The practical necessity of getting parents to assent to all experimental procedures prevents researchers from doing anything to an infant that is unpleasant. Moreover, in contrast to adult subjects, who are mostly readily available college students, infant subjects must be recruited from off campus in a process that is expensive and time-consuming. The high cost and relative rarity of infant subjects increases pressure to produce "good data," and yet experimenters have relatively little control over that outcome.

Options for both standardization and tacit knowledge are limited. Although standardization of the laboratory environment and procedures is achieved, the infants themselves remain highly idiosyncratic. Tacit knowledge certainly plays a role. Some experimentalists display skill in creating stimuli that children find interesting. Others are good at keeping them happy and calm. However, the nature of their experimental object limits the possibilities for tacit knowledge. Polanyi (1958) compared tacit scientific knowledge with the skills necessary to ride a bicycle. One can master riding a bike while being unable to explain it. However, although there are skilled infant researchers, infants cannot be "mastered." Even the most sensitive and brilliant researcher routinely faces uncooperative subjects.

To make matters more difficult for infant researchers, the discipline of psychology has long emphasized extreme rigorousness in methodological matters, an orientation

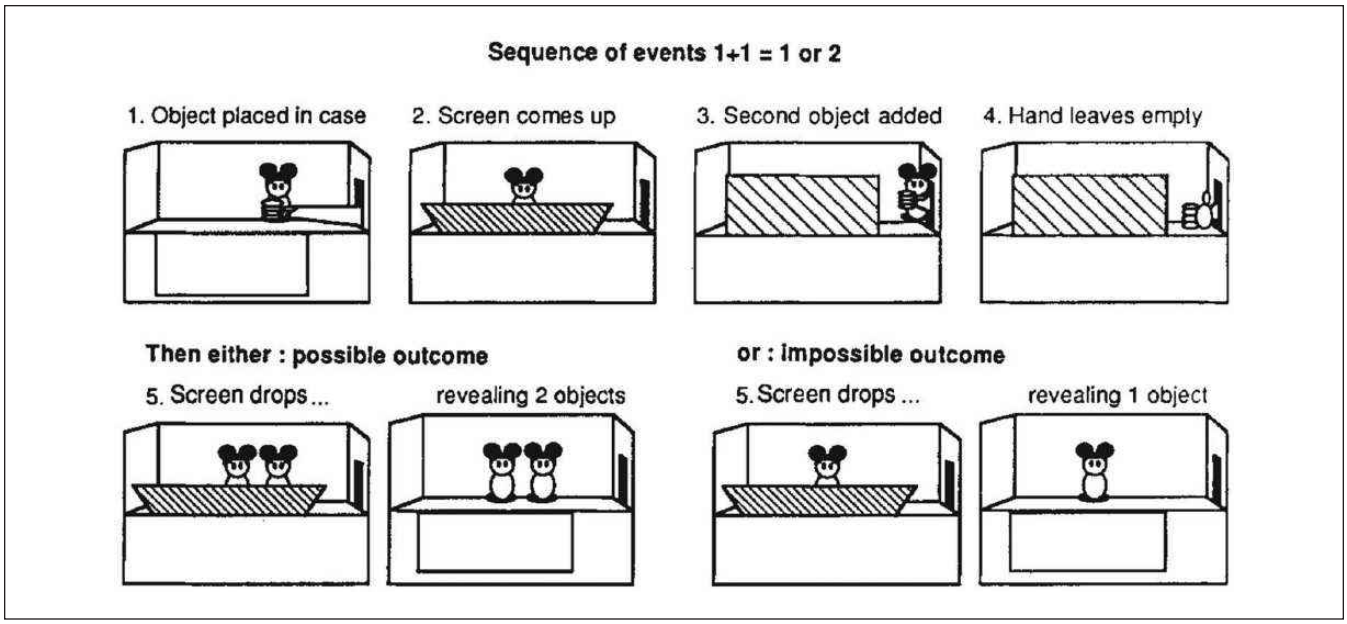

Figure 1. Addition and subtraction by human infants. Reproduced with permission from Wynn, K. 1992. "Addition and Subtraction by Human Infants." *Nature* 358(6389):749–50.

historian Kurt Danziger (1990) called "methodolotry." Furthermore, it is practically a requirement that findings meet a .05 level of statistical significance to be considered for publication (Cohen 1994; Porter 1995; Schmidt 1996). The scarce and recalcitrant object of study and the demands for rigorous methodology and statistical significance would seem to preclude a legitimate psychological science of infant cognition.

Yet infant cognition research is a popular and respected subfield of psychology. There are labs specializing in this research at nearly every top university, and developmental scientists routinely produce reports that appear in prestigious general science journals such as *Science* and *Nature*.

The Presentation of Infant Cognition Research

To illustrate what modern developmental science looks like, I will briefly summarize Karen Wynn's (1992) article "Addition and Subtraction by Human Infants." This is a highly esteemed article that has been cited more than 1,400 times. Its methodological and theoretical contributions influenced a number of the studies I discuss later. Sandwiched between a study of the earth's mantle and findings on gene interactions in drosophila in the journal *Nature*, the article's publication illustrates the prestigious position of modern developmental science.

In the article, Wynn (1992) argued that five-month-old infants display knowledge of simple arithmetic. In place of the observational methods used by early developmental psychologists, Wynn used the experimental methods that now dominate the field.

In one of the experiments (Figure 1), infants watched an object placed on a stage. A screen was then raised to obscure the object. The experimenter then, in view of the infant, placed a second object behind the screen. The screen was then lowered. In the control condition, there was the expected number of toys on the stage (two), whereas in the experimental condition, one of the toys was removed through a trapdoor and, thus, the stage contained a surprising number (one).

As in many infant cognition studies, infant expectation was inferred from the amount of time the subject looked at the stage. The belief underlying the method is that infants tend to look longer at things that are surprising or novel. Thus, because the 16 five-month-olds in the condition described looked longer at the mathematically impossible event for a statistically significant amount of time, Wynn (1992) concluded that infants have an innate knowledge of basic arithmetic operations, findings that have been replicated by multiple labs (Koechlin, Dehaene, and Mehler 1997; Simon, Hespos, and Rochat 1995).

Although an interesting study, this type of research begs the question, how are infants made into objects of study that can satisfy the methodological and statistical requirements of academic psychology? This requires an understanding of the lab environment and, specifically, the practices that do not make it into the published reports.

In a perfect experiment, the infant remains calm and focused throughout the process. There is a fine line between over- and understimulation. Tantrums and naps both derail experiments. Waiting rooms are designed specifically to keep children happy and engaged. They look like play rooms, with toys, crayons, and books. When the experiment begins, the parent (usually a mother) and child will go into a special

room designed for experiments. The parent will sit on a chair with the child in his or her lap, facing outward toward a wooden contraption that looks like a homemade puppet stage or, in more well funded labs, a large computer screen. The subject's attention must be directed toward the stimulus. Again, the design of the environment is meant to support this goal. The room will usually be dark, with the only light focused on the stage to direct the infant's attention. If the child remains calm and attentive, data can be gathered, usually by means of a video capture, which is either "live-coded" in a neighboring room via closed-circuit video or coded later from a recording. If the process goes smoothly, the infant stays relatively still, remains in frame, and does not allow shadows to obscure his or her eyes.

However, infant research rarely meets this ideal scenario. Infants are unpredictable, and researchers have few options for manipulating them. In the following sections, I present a different perspective on developmental research on the basis of my experience as a participant observer in three psychology laboratories specializing in infants and toddlers.

My goal is not simply to add to the literature highlighting the distance between published research and bench science. Instead, I will illustrate how experimental research on infants is characterized by a *specific set of problems* that have led researchers to adopt a series of mostly unstated solutions. Not mentioned in the published reports, these strategies are vital for maintaining productivity in the subfield but also makes the literature harder to interpret. To make my case, I conducted a 16-month ethnography at three developmental psychology laboratories (see Appendix A for additional information).

Strategies for Productivity in Developmental Psychology

This section outlines four strategies that aid the production of studies with statistical significance: protocol flexibility, stacking the deck, making experimental failures useful, and working backward from statistical significance.

Protocol Flexibility

Once in the lab, the experimenters attempt to collect useful data from children who are ignorant of the processes surrounding them. Tremendous effort is put into designing interesting, insightful studies. Lab meetings are valuable time, and many were spent entirely on the collective evaluation of study design. Moreover, there is an honest attempt to stay faithful to the protocols. Yet, the rarity of infant subjects, the unpredictability and ambiguity of their responses, and the pressures of the academic environment lead to situations in which experimenters are willing to overlook many of the hiccups that occur during experimentation.

Experimenters in developmental science use what might be referred to as a bend-but-don't-break philosophy of protocol adherence. The validity of experimental data may be conceived of as a continuum. At one end, there are data that come from an infant who is calm and focused, allowing the experimenter to capture the necessary data for proper analysis. At the other end are data that simply must be thrown out. This happens when the infant never stops crying or never shows any interest in the stimuli. Other times, the experimenter causes the failure. During one study, for instance, an experimenter hidden behind a stage was supposed to slowly turn a plastic barrel back and forth above the stage. Instead, she accidentally dropped it. The barrel landed on the stage with a thud and then rolled off the stage, crashing loudly on the floor. The data from that subject were thrown out. But between these two extremes lies an expansive gray area of minor violations that force researchers to decide whether to proceed or not.

Protocol violations tended to be context specific, yet there were several regular situations in which rules were bent. Some seemed relatively inconsequential. For instance, as a routine part of the experiments, parents are asked to close their eyes to prevent any unconscious influence on their children. Although this was explicitly stated in the instructions given to parents, during the actual experiment, it was often overlooked; the parents' eyes would remain open. Moreover, on several occasions, experimenters downplayed the importance of having one's eyes closed. One psychologist told a mother,

> During the trial, we ask you to close your eyes. That's just for the journals so we can say you weren't directing her attention. But you can peek if you want to. It's not a big deal. But there's not much to see.

Other violations had more potential to bias the data. For instance, studies were often stopped partway through so that the parent could change, feed, or calm down the infant. Later, the parent and child would reenter and begin again. Of course, at that point, the infant had already been exposed to some portion of the stimuli. Another time, an insistent older sibling demanded that he be allowed to join the mother and subject in the experiment room. During the trial, the mother had to tell the sibling to be quiet and still several times, which drew the attention of the infant away from the stage.

Because most of the props for stimuli are homemade and most of the people actually running the experiments are undergraduate assistants, it should not be surprising that experiments can be less than smooth. However, as long as problems did not completely derail the experiment, it went on. During one experiment I conducted, I was supposed to hide a ball in one of four buckets and allow the child to search for it. However, when I hid the ball it immediately rolled to the wall of the bucket and made an audible sound that both parent and child noticed. Instead of letting the child "search" for the ball at that point, I decided to hide the ball again. However, there was a precise order regarding where the ball was hidden in each trial, so I had to hide it in

the same place or risk making the data useless. After the trial, I told the experimenter who had been watching through the closed-circuit link what had happened, but she told me that the experiment went "great" and told me not to worry about it.

Outside of the experiment itself, the coding process was also full of ambiguity that was overcome through protocol violations. To measure looking time, the studies are coded by two "independent" coders (nearly always in the same room) on the basis of a closed-circuit video feed of the infant. However, there were many times when coding from the video became challenging. Very young infants would often fall in and out of frame because they lacked the strength to control their heads. Shadows made it difficult to read small eyes. It could be hard to tell whether the subject was actually looking at the stage or just staring off in that general direction. When two assistants coded, they often negotiated a joint solution to these problems. Half a dozen times while coding, I was asked by the other coder, "Is that looking?" "Are you marking this?" "Do you think that's a look?" During one study, an energetic subject refused to sit and, instead, stood on his mother's lap. Thus, we could only see the lower half of his face. The other coder told me to just code the chin. After the first few trials, the professor came in and asked, "Can you code from the chin?" The other coder replied, "Yeah, sort of," and we continued.

Although problems that arose during the studies were sometimes simply ignored, obvious breaches to procedure were noted on a sheet that the coder or experimenter filled out. One coder wrote "playing with shoes" next to the third and fourth trials on her sheet. Several experimenters wrote "fussy" to classify an entire experiment. But these data were not thrown out. The computer data output, sheets filled out by both experimenter and lead coder, and consent form were filed away together. Long after the experiment was run, when the data were analyzed, the researcher, who was often not present on the day of the experiment, decided whether the subject would be excluded because they were "distracted" or "fussy."

Stacking the Deck

In written reports, psychological experiments have a coherent narrative structure. A hypothesis is developed. Subjects are exposed to the stimulus and their responses are coded. When a predetermined number of subjects have been run, their data are then analyzed, and conclusions are drawn.

However, in actual lab practice, the experimental process is fluid. Instead of waiting for data from a set number of subjects to draw any conclusions, infant researchers have an ongoing relationship with data that begins as soon as the first subject is run. After coding a child who had just been run on a new experiment, a graduate student came into the coding room and asked to see the data. When she saw the computer printout, she began to jump up and down, squealing with joy. After she left, I asked the other coder why she was so happy. The coder explained that it was the first subject run in her new experiment, and the infant had responded as she had hoped. Although her reaction was unusually expressive, it is indicative of the relationship that experimenters have with data, even at very early stages in the experimentation process.

This was not unique to graduate students. I saw an undergraduate assistant come into her professor's office after video-coding an infant run in a new study. She told the professor that the effect was "12.4," which meant that the infant looked at the experimental condition for 12.4 seconds longer than the control condition. The professor then smiled and told me that 12.4 was a "huge effect." Before the undergraduate left, she told the professor that she had looked at the data from two other infants in that condition and both had shown similar effects.

Rather than waiting for the results from a set number of infants, experimenters began "eyeballing" the data as soon as babies were run and often began looking for statistical significance after just 5 or 10 subjects. During lab meetings and one-on-one discussions, experiments that were "in progress" and still collecting data were evaluated on the basis of these early results. When the preliminary data looked good, the test continued. When they showed ambiguous but significant results, the test usually continued. But when, after just a few subjects, no significance was found, the original protocol was abandoned and new variations were developed.

During one meeting, a psychologist was asking a postdoctoral researcher about his new experiment. It was not going well. The postdoc had run just three subjects and described the reactions of each in detail. One supported the hypothesis, one contradicted it, and one showed no preference for the experimental or control conditions. The professor responded, "Well, you can't tell from just three babies," but she gave him advice on how to alter the protocol slightly and instructed him to stop after 10 subjects if the study still was not working. In another meeting, the psychologist asked a new graduate student about a study. He told her he was reluctant to run statistics before all the data from all 16 subjects was in. She told him that if there was going to be an effect, it should be visible after 12 subjects, so he should run the statistics to find out.

Experimenters carefully attend to the computer printouts and run statistical tests long before they are finished collecting subjects. These serve as early signals that the experiment will be successful or a failure. Early signs of failure lead to adjustments so as not to waste time and resources. This makes sense from an economic standpoint. However, when a lab chooses to only complete the studies that show effects after a few subjects, they are essentially beginning each experiment with a head start. As the next section makes clear, however, this does not guarantee a successful study.

Experimental Failure Is Made into a Virtue

Papers in infant cognition often demonstrate some ability in a certain age group (e.g., 17-month-olds) and contrast it with failure from a younger age group (e.g., 14-month-olds). This can be used to demonstrate how knowledge develops in a particular domain. However, this need not be the order of actual research. In developmental psychology, failure often precedes success.

Because experiments on infant subjects are very costly in terms of both time and money, throwing away data is highly undesirable. Instead, when faced with a struggling experiment using a trusted experimental paradigm, experimenters would regularly run another study that had higher odds of success. This was accomplished by varying one aspect of the experiment, such as the age of the participants. For instance, when one experiment with 14-month-olds failed, the experimenter reran the same study with 18-month-olds, which then succeeded. Once a significant result was achieved, the failures were no longer valueless. They now represented a part of a larger story: "Eighteen-month-olds can achieve behavior X, but 14-month-olds cannot." Thus, the failed experiment becomes a boundary for the phenomenon.

In other cases, the experiment is simplified to increase the chances of a success. For instances, a postdoc and a research assistant were discussing an unsuccessful project with a professor. The study was modeled after the study by Wynn (1992) discussed above. Instead of using objects to test infants' knowledge of number, this experiment used images of people to test children's awareness of personal difference. If a cartoon man walks behind a screen and another who looks slightly different walks out, will the child expect the first man to still be there when the screen falls or do they lack the ability to differentiate people?

Unfortunately, the experiment had not yielded any significant results. The adviser told the postdoc and assistant that the stimuli were still too subtle for children this age. The cartoon people had to be extremely different. The research assistant suggested one be dressed in a hat and cape. The adviser jumped in, "And is black." The postdoc added, "And walks with a limp!" Although this was a tongue in cheek exchange, the meaning was clear when the adviser explained that they had to do whatever they could to get a successful test. She told them to "throw everything" at the babies in order to produce at least one experiment with statistical significance. The failures, she explained, could be framed around the success, as the limits to the phenomenon.

In another case, a graduate student was conducting an experiment modeled on a previously successful study from a psychologist from a different university. However, the experiment was not working, because the stimuli were boring, and most of the subjects were "fussing out." The psychologist told him, "It's important to interpret a failure in terms of a success" and suggested that he simplify his methods in order to achieve some significant result.

The strategy of finding virtue in failure is another economic decision to get as much utility as possible from the data (Collins 2003). If any success can be achieved, failures can be framed around it. One statistically significant finding can be the linchpin that holds a series of (mostly unsuccessful) studies together.

Working Backward from Statistical Significance

It is difficult to get statistically significant results when working with infants. However, it is even more difficult to get significant results that bear directly on the hypothesis that motivated the experiment. Often, statistically significant results present more questions than answers. Instead of conforming to the motivating hypothesis, the significant results are unpredicted, and their meaning is unclear. Roughly half of the regular lab meetings I attended (e.g., meetings concerned with research issues, not administration, planning, job searches, etc.) were dedicated to the discussion of statistically significant, but ambiguous, findings.

The structure of these meetings was similar across labs. A professor or graduate student would e-mail a short document to the lab a few days before and then hand out those same pages at the beginning of the meeting. Usually, they would contain a couple of box plots or bar charts. The experimenter would then point out where statistical significance was reached and then ask the lab for help figuring out what could be argued from the results. The lab would attempt to collectively craft a story out of the significant findings.

When a clear and interesting story could be told about significant findings, the original motivation was often abandoned. I attended a meeting between a graduate student and her mentor at which they were trying to decipher some results the student had just received. Their meaning was not at all clear, and the graduate student complained that she was having trouble remembering the motivation for the study in the first place. Her mentor responded, "You don't have to reconstruct your logic. You have the results now. If you can come up with an interpretation that works, that will motivate the hypothesis."

A blunt explanation of this strategy was given to me by an advanced graduate student: "You want to know how it works? We have a bunch of half-baked ideas. We run a bunch of experiments. Whatever data we get, we pretend that's what we were looking for." Rather than stay with the original, motivating hypothesis, researchers in developmental science learn to adjust to statistical significance. They then "fill out" the rest of the paper around this necessary core of psychological research.

As with protocol flexibility, there are limits to this sort of post hoc theorizing. During one meeting regarding a significant, but unclear, finding, a professor and graduate student went back and forth for 15 minutes discussing various hypotheses for the findings. Finally, the professor said, "I don't see a terrifically clear story coming from this," and

they moved on. In another case, a professor and research assistant were working on a grant application that contained some initial findings. One of the measures they were using was a composite of several tests. However, although the composite measure was significant, only one of the tests was driving the results. With it taken out, the composite measure was no longer significant. Unfortunately, the test was unrelated to the motivating hypothesis of the grant. For more than 20 minutes, they struggled to find a way to legitimize the composite measure. However, the professor decided that "it's a little dishonest to report the composite score if only [test A] is doing all the work." They decided to leave both the composite measure and the highly significant test out of the grant application.

Disciplinary Ideals and Local Culture

Psychology is currently in a period of methodological soul searching (John, Loewenstein, and Prelec 2012; Simmons, Nelson, and Simonsohn 2011; Wagenmakers et al. 2012). This is due to a string of recent events, including the publication of an article on precognition in a mainstream psychological journal (Bem 2011; Wagenmakers et al. 2011) and discoveries that a number of prominent psychologists manipulated or fabricated data (Carey 2011; Ferguson 2012; Wade 2010; Yong 2012).

Although incidents of outright fraud can be attributed to "a few bad apples," commentators have argued that a permissive attitude toward "questionable research practices" (Leahey 2008; Swazey, Anderson, and Lewis 1993) is a more pervasive and pernicious threat to psychology. One influential article argues that the validity of the psychological research literature has been undermined by unreported "researcher degrees of freedom," the decisions psychologists make during experiments and data analysis that are left out of published articles (Simmons et al. 2011). Another critical article highlights the

> uncomfortable fact that threatens the core of psychology's academic enterprise: almost without exception, psychologists do not commit themselves to a method of data analysis before they see the actual plan. It then becomes tempting to fine tune the analysis to the data in order to obtain a desired result—a procedure that invalidates the interpretation of common statistical tests. (Wagenmakers et al. 2012:632)

The problem with researcher freedom, according to these critiques, is that it casts doubt on the published literature. Simmons et al. (2011) argued that false positives are "perhaps the most costly error" because "once they appear in the literature, false positives are particularly persistent" (p .1). All of the strategies discussed above have dubious reputations among psychologists precisely because they increase the likelihood that false positives will enter the literature. A high volume of experiments (that are flexibly altered and abandoned) means that psychologists simply collect data until they begin to find statistical significance. Not every protocol violation will radically sway experimental outcomes, but some will. Building a story around significance does not always help enshrine a false positive, but sometimes it does.

Developmental scientists are well aware that there is a high risk for false positives in their field. Their relationship with published research is complex. Laboratories develop local knowledge regarding the validity or invalidity of articles, methods, and other labs on the basis of previous experience. Thus, claims become evaluated within a matrix of indicators. This is demonstrated by the internal use of replication within labs to evaluate published studies.

Labs will use methods innovated by outside researchers when moving into new research areas. However, when these experiments do not produce statistically significant results—when the infants are unable to sit through the experiment or when they show no awareness of the changing stimuli—it is not often clear why. To simplify somewhat, there are three possible hypotheses for failure. First, at the level of hypothesis testing, the extension of the experimental paradigm may simply show that the proposed relationship does not exist. For instance, an outside lab may have used a specific method to demonstrate an ability in six-month-olds. Trying the same experiment with four-month-olds may fail because the subjects are just too young and have not developed that ability. Second, however, there also may be deeper problems with the way the experiment was carried out. The new experiment may differ from its model in dozens of unintended ways and not be a "true" replication (Collins 1985). Third, and most problematically, the source study may simply be a false positive and thus impossible to replicate.

To sift through these competing explanations, psychologists who find their studies failing often conduct exact replications to test the method. When a graduate student was explaining the failure of a recent experiment to his adviser, she told him to replicate the original study with a few subjects: "We need a goddamn method check. The method has to work." This is a test for the first hypothesis. If the exact replication works, the experimenter may conclude that four-month-olds are simply too young to make the distinction asked of them in the new study. The children may truly be incapable at that age, or the stimuli may be too complex. Either way, the success of the replication provides some contrast for understanding the new experiment's failure.

If an exact replication fails, researchers begin a more thorough interrogation of the original study and its methods. Because developmental scientific reports present skeletal descriptions of the experiment, many aspects of the procedure are left out. Thus, when psychologists were having trouble getting a replication to work, they would call or e-mail the author of the source study to get a more detailed account of the experiment. In one case, the original author, a professor at a neighboring university, actually visited the lab

and watched as the experiment was being performed. It involved an experimenter manipulating objects on stage with a mechanical arm. She gave a series of instructions that were not in the original paper regarding how experimenters should pick up the objects and which way they should be looking during trials. The experiment still did not work, however, and was abandoned.

Copycatting methods leads to the growth of local knowledge regarding the validity and/or robustness of an article or line of research. If a source study came from a well-established lab yet could not be reproduced, the first thought is to assume fault. The experiment was treated as basically valid but difficult to reproduce. I heard these referred to as "fragile" paradigms.

However, when the author is relatively unknown or the experiment still does not work after several attempts, the original study becomes marked as dubious in the lab. During one conversation, a graduate student was discussing an article from an unknown lab that pertained to her project. Her adviser dismissed the article because "no one's been able to reproduce it." Because negative findings are rarely published, this knowledge does not diffuse across the field through the medium of journal articles. Instead, failures become known within the lab and across labs through interpersonal networks.

The Baby and the Bathwater

In developmental science, researchers face a tricky balancing act. On one side, there are psychology's inflexible ideals regarding what constitutes good work. By modeling their field on an idealized conception of scientific research in the natural sciences, psychologists have produced an unforgiving culture in which experimental designs must be flawless and statistical significance must be achieved. On the other side, however, is a room full of infant subjects perpetually riding the razor's edge between a stormy tantrum and a sound sleep.

Working with infants demands that researchers frequently use local, contingent decision making. Of course, such decisions are at the heart of expert judgment (Daston and Galison 2010) and do not necessarily signal unethical or problematic practice. Much of the literature in the area of laboratory ethnography has demonstrated how researchers struggle to control variability and produce stable effects. However, what distinguishes developmental psychology from the topics of previous studies—mainly concerned with biology and physics labs—is that, for those who research infants, there is little hope of ever reaching "interactive stabilization" (Pickering 1995) between researcher, research technology, and research object. Neither improvements in technology nor more embodied skill will make an infant controllable (Peterson 2015), and this has a significant effect on how the field develops.

I conclude by arguing that fields can meet the challenge of difficult research objects in one of three ways. First, the realm of questions can be narrowed to match the capabilities of the research object. Second, standards of rigor can be loosened to allow researchers to maintain productivity while acknowledging the general diminution of probative value of individual studies. Finally, researchers can engage in "questionable research practices" in order to meet disciplinary standards (see also Roth and Bowen's [2001] discussion of "creative solutions" and "fibbing" by field ecologists). Although these practices are widespread in developmental psychology, they are becoming increasingly unacceptable. Commentators have argued that they contribute to the nonreplicability of the field and, thus, undermine the long-term health of the discipline.

However, there are reasons to be skeptical of the Manichaean division between "good" research that adheres to the strictures of hypothesis testing and "bad" research that introduces a higher probability for false positives. Although these questions are often framed in terms of scientific ethics, different epistemic cultures may adopt different research standards that are each defensible. For instance, Collins (1998) illustrated how communities of physicists could be distinguished on the basis of how much data processing they believed was necessary prior to publication. Researchers in "open evidential cultures" were willing publish relatively unprocessed data into print. In their opinion, the risk for error was outweighed by the potential for major advances. Moreover, they believed the wider community of scientists would help separate the wheat from the chaff through independent replications. In contrast, those in "closed evidential cultures" believed publication should be reserved for highly processed data in order to reduce the amount of error in the literature.

Open evidential cultures may be defensible under certain conditions. When problems are pressing and progress needs to be made quickly, creativity may be prized over ascetic rigor. Certain areas of medical or environmental science may meet this criterion. Developmental psychology does not. However, it may meet a second criterion. When research findings are not tightly coupled with some piece of material or social technology—that is, when the "consumers" of such science do not significantly depend on the veracity of individual articles—then local culture can function as an internal mechanism for evaluation in the field. Similar to the way oncologists use a "web of trials" rather than relying on a single, authoritative study (Keating and Cambrosio 2011) or how weather forecasters use multiple streams of evidence and personal experience to craft a prediction (Daipha 2010; Fine 2007), knowledge in such fields may develop positively even in a literature that contains more false positives than would be expected by chance alone.

Appendix A

Methods

Over 16 months, I was a participant observer in three psychology laboratories that specialize in psychological experimentation on infants and toddlers. Two of the labs were at a

private university in the Midwest, and one was at a prestigious private university on the East Coast. In one of the labs, I attended weekly lab meetings for a semester. In the second, I volunteered with the day-to-day running of the lab one to two days a week for one year. In the last, I worked in the lab between 35 and 40 hours a week for five weeks.

The labs varied in size. The smallest housed only a faculty member, a single graduate student, and 4 to 6 transitory undergraduate assistants. The largest lab was the bustling home of a faculty member, 2 postdoctoral researchers, 2 independently supported research fellows, 2 salaried lab managers, 8 graduate students, and 5 to 10 undergraduate research volunteers. One of the labs could be classified as small, and two were large.

The routine of laboratory life was largely consistent across sites. Undergraduates were in charge of scheduling subjects and often participated in the experimentation process as either coders who watched live or recorded video of the experiment in order to code subject responses or, sometimes, as experimenters (i.e., actually dealing with the child subjects). However, graduate students were usually the experimenters for their own experiments. In the smaller labs, the faculty members were involved with the experimentation process while, in the larger labs, they focused more on the intellectual and administrative aspects of running a large lab. This left the routine aspects of experimentation to graduate students, research fellows, and postdocs. During most weeks, labs would gather data on 5 to 15 different experiments.

There were regularly scheduled lab meetings that allowed the entire lab to congregate and discuss ongoing work. During these meetings, one of the lab members would present either an idea for research or early results from an ongoing experiment to solicit advice and expose fragile work to friendly criticism.

Although my role was different in each lab, resulting in differential access, across all labs I took part in nearly every aspect of laboratory life. This included recruiting and scheduling subjects, updating databases, training undergraduate assistants, conducting and coding experiments, observing both laboratory-wide and smaller project-based meetings, and participating in theoretical discussions.

Notes were taken throughout the course of the day. All direct quotations were written down immediately. Field notes were coded inductively on ATLAS.ti.

References

Bem, D. J. 2011. "Feeling the Future: Experimental Evidence for Anomalous Retroactive Influences on Cognition and Affect." *Journal of Personality and Social Psychology* 100(3):407–25.

Carey, B. 2011. "Fraud Case Seen as a Red Flag for Psychology Research." *The New York Times*. Retrieved October 17, 2012 (http://www.nytimes.com/2011/11/03/health/research/noted-dutch-psychologist-stapel-accused-of-research-fraud.html?_r=0).

Cohen, J. 1994. "The Earth is Round ($p<.05$)." *American Psychologist* 49(12):997–1003.

Collins, H. M. 1974. "The TEA Set: Tacit Knowledge and Scientific Networks." *Science Studies* 4:165–86.

Collins, H. M. 1985. *Changing Order: Replication and Induction in Scientific Practice*. London: Sage Ltd.

Collins, H. M. 1998. "The Meaning of Data: Open and Closed Evidential Cultures in the Search for Gravitational Waves." *American Journal of Sociology* 104(2):293–338.

Collins, H. M. 2001. "Tacit Knowledge, Trust and the Q of Sapphire." *Social Studies of Science* 31(1):71–85.

Collins, H. M. 2003. "Lead into Gold: The Science of Finding Nothing." *Studies in History and Philosophy of Science: Part A* 34(4):661–91.

Collins, H. M. 2010. *Tacit and Explicit Knowledge*. Chicago: University of Chicago Press.

Daipha, P. 2010. "Visual Perception at Work: Lessons from the World of Meteorology." *Poetics* 38(2):151–65.

Danziger, Kurt. 1990. *Constructing the Subject: Historical Origins of Psychological Research*. Cambridge, UK: Cambridge University Press.

Daston, L. and Galison, P. 1992. "The Image of Objectivity." *Representations* 40(1):81–128.

Daston, L. and Galison, P. 2010. *Objectivity*. New York: Zone.

Delamont, S. and Atkinson, P. 2001. "Doctoring Uncertainty: Mastering Craft Knowledge." *Social Studies of Science* 31(1):87–107.

Epstein, S. 2007. *Inclusion: The Politics of Difference in Medical Research*. Chicago: University of Chicago Press.

Ferguson, C. J. 2012. "Can We Trust Psychological Research?" *Time*. Retrieved October 17, 2012 (http://ideas.time.com/2012/07/17/can-we-trust-psychological-research/).

Fine, G. A. 2007. *Authors of the Storm: Meteorologists and the Culture of Prediction*. Chicago: University of Chicago Press.

Fujimura, J. H. 1992. "Crafting Science: Standardized Packages, Boundary Objects, and 'Translation.'" Pp. 168–211 in *Science as Practice and Culture*, edited by A. Pickering. Chicago: University of Chicago Press.

Gilbert, G. N. and Mulkay, M. 1984. *Opening Pandora's Box: A Sociological Analysis of Scientists' Discourse*. Cambridge, UK: Cambridge University Press.

Greenwood, J. D. 1999. "Understanding the 'Cognitive Revolution' in Psychology." *Journal of the History of the Behavioral Sciences* 35(1):1–22.

Holton, G. 1978. "Subelectrons, Presuppositions, and the Millikan-Ehrenhaft Dispute." *Historical Studies in the Physical Sciences* 9:161–224.

John, L. K., Loewenstein, G., and Prelec, D. 2012. "Measuring the Prevalence of Questionable Research Practices with Incentives for Truth Telling." *Psychological Science* 23(5), 524–32.

Jordan, K. and Lynch, M. 1998. "The Dissemination, Standardization, and Routinization of a Molecular Biological Technique." *Social Studies of Science* 28(5/6):773–800.

Keating, P. and Cambrosio, A. 2011. *Cancer on Trial: Oncology as a New Style of Practice*. Chicago: University of Chicago Press.

Knorr-Cetina, K. 1983. "The Ethnographic Study of Scientific Work: Towards a Constructivist Interpretation of Science." Pp. 115–40 in *Science Observed: Perspectives on the Social Studies of Science*, edited by K. Knorr-Centina. London: Sage Ltd.

Knorr Cetina, Karen. 1995. "Laboratory Studies: The Cultural Approach to the Study of Science." Pp. 140–67 in *The Handbook of Science and Technology Studies*, edited by S. Jasanoff, G. E. Markle, J. C. Petersen, and T. Pinch. Thousand Oaks, CA: Sage.

Koechlin, E., Dehaene, S., and Mehler, J. 1998. "Numerical Transformations in Five-month Old Infants." *Mathematical Cognition* 3(1):89–104.

Kohler, R. E. 1999. "Moral Economy, Material Culture, and Community in Drosophila Genetics." Pp. 243–57 in *The Science Studies Reader*, edited by M. Biagioli. New York: Routledge.

Latour, B. 1983. "Give Me a Laboratory and I Will Raise the World." Pp. 141–69 in *Science Observed: Perspectives on the Social Study of Science*, edited by K. Knorr-Cetina and M. Mulkay. London: Sage Ltd.

Leahey, E. 2008. "Overseeing Research Practice: The Case of Data Editing." *Science, Technology and Human Values* 33(5): 605–30.

Lemov, R. 2005. *World as Laboratory: Experiments with Mice, Mazes, and Men*. New York: Hill & Wang.

Lynch, Michael. 1985. "Discipline and the Material Form of Images: An Analysis of Scientific Visibility." *Social Studies of Science*, 15(1), 37–66.

MacKenzie, D. and Spinardi, G. 1995. "Tacit Knowledge, Weapons Design, and the Uninvention of Nuclear Weapons." *American Journal of Sociology* 101(1):44–99.

Milgram, S. 1963. "Behavioral Study of Obedience." *Journal of Abnormal and Social Psychology* 67(4):371–8.

Peterson, D. 2015. "All That Is Solid: Bench-building at the Frontiers of Two Experimental Sciences." *American Sociological Review* 80(6):1201–25.

Pickering, A. 1995. *The Mangle of Practice: Time, Agency, and Science*. Chicago: Chicago University Press.

Polanyi, M. 1958. *Personal Knowledge: Towards a Post-critical Philosophy*. Chicago: University of Chicago Press.

Porter, T. M. 1995. *Trust in Numbers: The Pursuit of Objectivity in Science and Public Life*. Princeton, NJ: Princeton University Press.

Roth, W. M. and Bowen, G. M. 2001. "'Creative Solutions' and 'Fibbing Results': Enculturation in Field Ecology." *Social Studies of Science* 31(4):533–56.

Schmidt, F. L. 1996. "Statistical Significance Testing and Cumulative Knowledge in Psychology: Implications for Training of Researchers." *Psychological Methods* 1(2):115–29.

Simmons, J. P., Nelson, L. D., and Simonsohn, U. "False-positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant." *Psychological Science* 22(11):1359–66.

Simon, T. J., Hespos, S. J., and Rochat, P. 1995. "Do Infants Understand Simple Arithmetic? A Replication of Wynn 1992." *Cognitive Development* 10(2):253–69.

Skinner, B. F. 1938. *The Behavior of Organisms*. New York: Appleton-Century Crofts.

Swazey, J. P., Anderson, M. S., and Lewis, K. S. 1993. "Ethical Problems in Academic Research." *American Scientist* 81(6):542–53.

Timmermans, S. and Epstein, S. 2010. "A World of Standards but Not a Standard World: The Sociology of Standards and Standardization." *Annual Review of Sociology* 36:69–89.

Wade, N. 2010. "Harvard Finds Scientist Guilty of Misconduct." *The New York Times*. Retrieved October 17, 2012 (http://www.nytimes.com/2010/08/21/education/21harvard.html).

Wagenmakers, E. J., Wetzels, R., Borsboom, D., and van der Maas, H. L. J. 2011. "Why Psychologists Must Change the Way They Analyze Their Data: The Case of Psi: Comment on Bem (2011)." *Journal of Personality and Social Psychology* 100(3):426–32.

Wagenmakers, E., Wetzels, R., Borsboom, D., van der Maas, H., and Kievit, R. 2012. "An Agenda for Purely Confirmatory Research." *Perspectives on Psychological Science* 7(6): 632–38.

Woolgar, S. 1982. "Laboratory Studies: A Comment on the State of the Art." *Social Studies of Science* 12(4):481–98.

Wynn, K. 1992. "Addition and Subtraction by Human Infants." *Nature* 358(6389):749–50.

Yong, E. 2012. "Replication Studies: Bad Copy." *Nature* 485(7398):298–300.

Author Biography

David Peterson is a PhD candidate in the sociology program at Northwestern University. His interests include social theory, the sociology of science and knowledge, and economic sociology. His dissertation investigates competing ideas of scientific progress through interviews and four years of ethnographic observations in 12 psychology and neuroscience labs. His previous research has been published in the *American Sociological Review, Symbolic Interaction*, and other outlets. He is currently looking at the moral aspects of scientific markets through a case study of recent scandals in social psychology.