CONCEPTUAL PROBLEMS IN STATISTICS, TESTING AND EXPERIMENTATION

Chapter for *Routledge Companion to the Philosophy of Psychology*

David Danks[a,b]

Frederick Eberhardt[a]

[a] – Department of Philosophy, Carnegie Mellon University

[b] –Institute for Human & Machine Cognition

Direct correspondence to:

David Danks
Department of Philosophy
135 Baker Hall
Carnegie Mellon University
Pittsburgh, PA 15213
ddanks@cmu.edu
Tel: (412) 268-8047
Fax: (412) 268-1440

Introduction

By virtue of the individuals and processes being studied, cognitive psychology is a methodologically challenging science. On the one hand, psychology aims to be a natural science, complete with testing of models by experimental evidence. On the other hand, the distinctive domain that psychology considers cannot be studied with many standard experimental techniques. This chapter explores some of the methodological and statistical challenges confronting present-day psychology, with a principal focus on problems that are particular to psychology, as opposed to the general problems of experimental design, confirmation, inference, and so on that confront essentially all sciences. We focus on problems for which no full and complete solution is known. Where appropriate, we have indicated various possibilities or proposals, but many of these challenges remain open questions for the practice of psychology.

There is one important body of difficulties that we will not explore in detail: namely, the relative inaccessibility of the target phenomena of psychology. Psychological objects such as beliefs and desires are not directly accessible to experimenters, either for manipulation or measurement. We cannot directly intervene on psychological states; at most, we can provide stimuli designed to induce a particular mental state, though those interventions are rarely as precise as desired (Campbell, in press). We thus confront a host of methodological concerns (e.g., Do our manipulations affect their target, but not other causally relevant variables?) and conceptual concerns (e.g., Are mental states causal variables at all? Does multiple realizability (if correct) preclude the very possibility of unambiguous interventions?). The lack of direct accessibility also presents a measurement challenge: we cannot directly observe mental states, but rather must measure various proxies, typically readily observable behavior. We note this measurement difficulty not as any sort of call for behaviorism; inference about unobserved

entities is difficult, but certainly not impossible. Rather, our inability to directly observe mental states places an additional methodological burden on the psychologist: she must ensure that proxy measurements are suitably correlated with the mental states that she is trying to investigate. As a practical example, one should be skeptical when interpreting data obtained from surveys, as they are typically only (very) noisy proxies for the relevant underlying mental states. Furthermore, the measurement and manipulation challenges interact: if we could directly measure mental states, then our inability to directly manipulate them would pose less of a problem; if we could directly manipulate mental states, then we could use those manipulations to help directly measure them.

These two challenges are significant, but have also been the target of many different philosophical debates (e.g., about the nature of mental causation). In contrast, we focus on conceptual and methodological challenges that are more specific to the practice of contemporary psychology, and that have either not received as much attention, or continue to be the subject of methodological debate. In particular, our discussion will center on issues raised by the significant variability found in psychological data, the challenges presented by unconscious cognition, and important methodological issues in experiment design and analysis.

Variability in Psychological Data

One of the most noticeable features of psychological data – particularly to newcomers to the field – is the combination of large variability in the data and (relatively) small sample sizes. In general, we have neither sufficiently clean data to make precise predictions, nor sufficient sample sizes to overcome data noise from many different sources. As a result, we must accept the fact that theoretically distinguishable theories will sometimes not be distinguishable given the data at hand, and may not be distinguishable given *any* plausible future data. For example,

exemplar-based and prototype-based theories of categorization make differing predictions for certain experimental designs, but experiments using those designs do not discriminate between the theories. Superior performance by one or the other model-type is almost certainly due to superior fitting of noise in people's response mechanisms, rather than underlying features of their categorization mechanisms (Minda & Smith, 2001; Nosofsky & Zaki, 2002; Olsson, Wennerholm, & Lyxzèn, 2004). Of course, all sciences face the noise vs. sample size tradeoff in one form or another; psychology is not special in this regard, though the problem is more extreme here. Psychology also has distinctive *reasons* for the variability. In general, we can think about this variability as arising from three distinct sources: (1) participant interest; (2) individual differences; and possibly (3) stochastic mechanisms. In this section, we examine conceptual issues that arise from each of these sources of variability, as well as their interaction.

A persistent (though rarely discussed) challenge of psychological experimentation is the obvious fact that the data points come from intentional agents with their own desires, beliefs, etc. In particular, there is a general, unstated recognition that some subset of the experimental participants will either (i) fail to understand the experiment instructions, or (ii) not be sufficiently motivated to actually follow the instructions. This problem is arguably exacerbated by the common failure to check whether an experiment provides an appropriate incentive structure to elicit a participant's 'best' behavior, as well as the widespread use of undergraduate students as experimental participants. The prototypical 'problem participant' here is one who answers 'No' or '0' for every question in an effort to leave as quickly as possible. A more difficult case is a participant in an experiment on arithmetic ability who aims to finish quickly by providing approximately correct answers, such as '$101 \times 59 = 5900$.' These responses do not accurately measure the participant's ability, and so should presumably be excluded from our analyses and

theorizing. But what distinguishes this individual from the participant who simply is not very good at arithmetic?

Various experimental techniques can mitigate this problem, but it is rarely possible to eliminate it completely for interesting experiments. Instead, we need a method for classifying some participants as 'unresponsive' or 'failing to follow instructions.' Such a method will necessarily depend on a normative model of how people *ought* to behave in this experimental setting (e.g., that no one should ever say '$101 \times 59 = 0$'). The problem is that normative models of higher-level cognition should be sensitive to the particular limits and constraints of our cognitive system, at least if one believes that normative models should not require behavior that is in principle unattainable (i.e., if one thinks that 'ought implies can'). We thus face a potential conceptual circle: namely, our normative models of human cognition should be sensitive to descriptive capacities, but to develop models of descriptive capacities, we use the same normative models to interpret and identify relevant experimental data (see also Harré, 1996).[1] Other sciences typically avoid this circle by finding an independent ground for the normative model in its more general applicability (e.g., using Newtonian mechanics to predict the specific behavior of a pendulum); it is not clear that normative models in psychology have the necessary independent grounding, or that there exist framework theories with the appropriate generality.

This circularity need not be a vicious one. There are at least two alternatives, though neither has been carefully explored. First, one could provide a bootstrap account in which descriptive accounts of cognitive phenomena at a 'lower' level (e.g., short-term memory) are used to justify normative theories of 'higher' phenomena (e.g., causal learning), which are then

[1] Note that this problem is distinct from various debates in psychology about which normative model is appropriate for a particular experimental design (e.g., predicate logic vs. Bayesian hypothesis confirmation in the Wason selection task).

used to inform the development of descriptive theories at that same higher level. This approach requires an independently justified 'ground level' theory (e.g., a descriptive account of working memory based on neuroscientific data), as well as evidence that the bootstrapping could actually work in practice. A second response to the circularity would aim for equilibrium between the various theories: the data deemed 'acceptable' by the normative model *N* at one level would support a descriptive model *D* at the same level whose constraints are consistent with *N*. An equilibrium-based account would then need to establish a connection between this type of equilibrium and the truth about cognitive mechanisms, and it is not obvious that any such connection must hold. In practice, psychology almost certainly involves a mix of responses to the general conceptual circle. Regardless of approach, though, we need a more rigorous understanding of the relationship between normative and descriptive theories of cognition (see also Colyvan, in press).

Even if we solve this general problem of 'outlier' detection, the other sources of variability complicate discovery and inference from psychological data. Most of cognitive psychology is individualistic in target: it aims to develop models of cognitive functioning in particular individuals, for purposes of prediction, explanation, and understanding. At the same time, the significant variability of even 'clean' data naturally leads to the use of population-level statistics (e.g., mean rating) for testing experiments. Since data about the whole group of participants is more robust against some of the problems of (random) noise, psychologists frequently use it to make inferences about features of the cognitive mechanisms in a particular individual. There are, however, serious challenges in making inferences from features of the population to features of the individual. For example, some population-level statistics provably

do not match the individual-level statistics, even if every individual in the population is identical (Chu, Glymour, Scheines, & Spirtes, 2003; Danks & Glymour, 2001).

Theories are also frequently tested by checking model fit, but the best-fitting model for populations need not be the same as the best-fitting model for individuals, either in theory (Brown & Heathcote, 2003; Myung, Kim, & Pitt, 2000) or in practice. For example, participants in many categorization experiments learn novel categories, provide ratings (e.g., the likelihood of some new object being an *A*), and then those mean ratings are compared to the predictions of various theories. One of the most commonly used category structures is the so-called 5/4 structure (Medin & Schaffer, 1978), and exemplar-based theories provide the best model first for the mean ratings of most experiments with this structure. In contrast, Minda & Smith (2002) argue that prototype-based models provide a better fit for each *individual* (though see Zaki, Nosofsky, Stanton, & Cohen, 2003). Thus, the best-fitting categorization model for the population might not be the best model for each of the individuals.

Moreover, simple uses of population-level statistics to make inferences about individuals require an assumption of population uniformity, and there are numerous cautionary tales of empirical failures of this assumption. As just one example, in a standard type of causal learning experiment, participants observe a sequence of cases and then provide judgments about the causal strength of one factor to bring about another. The psychologist is principally interested in the way judgments vary as a function of the statistics of the sequence, and so the standard data analysis is to compare the mean ratings of causal efficacy against various theories. However, multiple studies provide evidence that participants are actually using a range of strategies (Anderson & Sheu, 1995; Buehner, Cheng, & Clifford, 2003; Lober & Shanks, 2000), and so mean ratings are not necessarily informative about features of the individual learners. And the

method of analysis matters for the conclusion: a single-group reanalysis of data from Buehner & Cheng (1997) supported Bayesian causal support (Griffiths & Tenenbaum, 2005), while a two-group reanalysis supported a mixture of causal power and conditional ΔP (Buehner*, et al.*, 2003).

The standard response to all of these worries is to pursue so-called individual differences research. There are two distinct types of individual differences research: one in which all individuals are assumed to have the same algorithm, but different parameter values; and a second in which individuals are allowed to have different underlying algorithms. Research in the first vein has a long history in psychometrics (including seminal work such as Spearman, 1904), and assumes that differential performance arises because people differ in some underlying attribute that plays an important role in cognitive function. The approach is analogous to understanding differential performance as arising from the same program running on two computers that have hardware differences (e.g., processor speed). This line of research has typically focused on 'important' individual-specific parameters, such as working memory capacity, processing speed, or general intelligence, as well as their interaction (e.g., Ackerman, 1988; Barrett, Tugade, & Engle, 2004; Just & Carpenter, 1992; Schunn, Lovett, & Reder, 2001; Sternberg, 1977; Taatgen, 2002). The standard experimental design for this type of research is to provide people with a battery of psychometric tests to estimate the relevant global parameter(s), and then to correlate those estimates with performance on a subsequent, more complicated task. Importantly, this research assumes that people all use fundamentally the same algorithm or underlying cognitive mechanism; they just differ in their ability to carry out the process.

In contrast, the second type of individual differences research allows for the possibility that people use completely different algorithms to solve problems. There is a significant body of research in a wide range of domains and for many different populations showing that people use

a variety of strategies (e.g., Chi, Feltovich, & Glaser, 1981; Ericsson & Charness, 1994; Payne, Bettman, & Johnson, 1993; Siegler, 1996; Stanovich, 1999). This type of individual differences research uses an individual's behavior pattern to determine which strategy she is most likely using and, if relevant, the parameters for that algorithm (e.g., Lee, 2006; Schunn & Reder, 1998). Continuing the analogy with computers, this type of analysis models differential performance as resulting from the same computer (or same hardware) running different programs that achieve similar goals (e.g., Matlab vs. Excel).

Regardless of focus, individual difference analyses must provide a model for each individual participant. Given the relative noisiness of psychological data, one typically must collect many more data points for an individual differences analysis than for a more traditional population-level analysis. Moreover, modeling each individual separately can make it difficult to balance both goodness of fit and generalizability for the model, as it becomes even more difficult to know which variations are due to noise, and which reflect actual features of the underlying cognitive mechanisms (Pitt, Myung, & Zhang, 2002). Thus, a promising line of recent research assumes that there is some small number of distinct groups, where every individual within a group has the same algorithm and parameters. Since groups have multiple members, one gains the benefits of multiple measurements to offset noise; at the same time, one is not locked into the assumption that every individual is identical. The tricky part is, of course, determining the appropriate number of groups, as well as which participant belongs to which group. There are a number of sophisticated statistical techniques that have recently been proposed, including Dirichlet process models (Navarro, Griffiths, Steyvers, & Lee, 2006), a hierarchical Bayesian framework (Rouder, Sun, Speckman, Lu, & Zhou, 2003), or maximum likelihood partition generation (Lee & Webb, 2005). At their core, they are all structurally similar: find (i) the

number of groups, (ii) allocation of individuals to groups, and (iii) characteristics of each group that maximizes the likelihood of observing data such as this. As such, they are typically quite computationally complex, but have the potential to establish a middle ground between population-level and individual differences analyses.

Finally, the possibility of stochastic cognitive mechanisms raises important questions about what types of predictions we ought to expect from psychological models. Deterministic psychological models are straightforward to confirm, at least from a theoretical point-of-view. Since precise predictions are made for each individual and each situation, we can generate the proper predictions (perhaps with noise) and check whether the observed data matches those predictions (up to noise). In contrast, models that posit stochastic cognitive mechanisms cannot be confirmed in the same manner, since there is no determinate prediction for each situation. As just one example of many, the category learning model RULEX (Nosofsky & Palmeri, 1998; Nosofsky, Palmeri, & McKinley, 1994) includes stochastic choice points during learning. If the same individual saw the exact same sequence of data on two separate occasions, she would not necessarily learn the same category, and so RULEX predicts only a distribution of learned categories for any individual. Thus, one is almost forced to confirm the RULEX model – or any other model that posits inherently stochastic elements – by comparing a predicted distribution of responses against the population as a whole, particularly since one can only provide an experimental participant with the same sequence multiple times by changing experimental features that might be relevant to their learning (e.g., cover story). Individual difference analyses are (almost) ruled out from the start, and even group-level analyses are more difficult.

The possibility of stochastic cognitive mechanisms raises one final issue: namely, the fact that we rarely are able to test psychological theories in isolation. The worry is not a general,

Quine-Duhem one about the testability of theories, but rather that almost any theory of learning or reasoning must be supplemented by a choice or response theory in order to make predictions about observable behavior. Consider a standard categorization problem in which a psychological model might predict that P(Object X is in Category A) = 0.75 and P(Object X is in Category B) = 0.25, but where the experimental participant must make a forced-choice of X as definitely an A or a B. The categorization theory alone does not make any particular prediction about the participant's response; rather, one must also state how the participant's beliefs are converted into a response. The standard choice model (assumed without discussion in many cases) is probability matching: the probability of any response is equal to the probability of that possibility. In the example above, probability matching predicts P(Respond 'A') = 0.75 and P(Respond 'B') = 0.25. When the assumption of probability matching has actually been tested, however, it has not fared well, at least in the domain of categorization (Ashby & Gott, 1988; Nosofsky & Zaki, 2002; Wills, Reimers, Stewart, Suret, & McLaren, 2000). Thus, stochastic mechanisms introduce yet another layer of model testing to psychology.

Implicit vs. Explicit Mental States

Although no mental states are, at one level, directly observable, we nonetheless think that we can frequently obtain close-to-direct measurements of them. In particular, many experiments in cognitive psychology assume that verbal reports (e.g., ratings on a seven-point Likert scale, descriptions of current beliefs, etc.) give more-or-less direct access to the relevant mental states. The discussion in the previous section implicitly made precisely this type of assumption. At the same time, clearly not all mental states can be reported in this manner. Some behavior is generated by unconscious cognitive mechanisms about which we cannot give verbal reports. These unconscious processes can be quite simple and mundane; we are not referring here to the

elements of, say, Freudian psychology. For example, one early experiment on so-called implicit learning (Reber, 1967) asked participants to learn letter sequences based on an unknown-to-the-participant artificial grammar. Participants were subsequently able to distinguish letter sequences satisfying the artificial grammar with better-than-chance accuracy, even though they were completely unable to articulate the method by which they performed the classification. The general tactic of finding significant differences in behavior without any corresponding reports of awareness has led to an astonishingly large body of research on unconscious processing, including implicit learning (Goschke, 1997; Shanks, 2005, provide reviews), implicit memory (Schacter, 1987; Schacter, Chiu, & Ochsner, 1993, provide reviews), and other types of implicit cognition (e.g., Dienes & Perner, 1999; Fazio & Olson, 2003; Underwood, 1996).

Implicit cognition is variously understood roughly as cognition that either has no accompanying awareness, or is not affected by lack of awareness. The central methodological challenge is thus to use either reliable measures or careful experimental design to determine whether some cognition occurs with absolutely no awareness. While careful experimental design is important, it seems highly unlikely that any experimental design can *guarantee* that participants will have no awareness of the relevant information. With regards to the first possibility, an open research question is precisely whether various subjective and objective measures are reliable indicators of *actual* lack of awareness (e.g., Dienes, Altmann, Kwan, & Goode, 1995; Dienes & Scott, 2005; Tunney & Shanks, 2003). Not surprisingly, many subjective measures, such as introspective reports or the ability to explain one's own behavior, are not reliable measures of level of awareness (Shanks, 2005). In contrast, objective measures of the implicitness of some cognition aim for a behavioral measure of lack of awareness (e.g., inability to use some information for a basic forced-choice task), but where the information affects their

behavior in more subtle ways. Perhaps the best-known example of such an effect is the 'mere exposure effect' (Kunst-Wilson & Zajonc, 1980; Mandler, Nakamura, & van Zandt, 1987). In the classic version of this experiment, participants were briefly shown various geometric figures, and then later shown each figure with a novel geometric figure and asked to choose which figure they recognized, and which they liked more. Participants performed at chance in terms of recognizing which figure of the pair had previously occurred, but were significantly more likely to 'like' the previously observed figure. Such effects have been shown to last for as long as 17 years (Mitchell, 2006). The standard interpretation is that participants have no explicit memory of seeing the figure, but do have some type of implicit memory of the previous exposure, which then influences their 'liking.' Alternative explanations for these findings (e.g., Whittlesea & Price, 2001) do not posit lack of awareness, though, and so the possibility of objective measures remains an open question.

More generally, the possibility of implicit cognition raises serious methodological concerns about the common psychological practice of using verbal reports as a major source of experimental data. If significant elements of our cognition occur without any corresponding awareness, then the assumption that verbal reports provide reliable measurements of underlying mental states might be less justified than is typically thought. Similar concerns are raised by experiments demonstrating a range of metacognitive failures, including evidence that people often act for reasons of which they are unaware (Kruger & Dunning, 1999; Nisbett & Ross, 1991). One response to these concerns has been a shift towards experiments that are grounded in behavioral measures, rather than verbal ones. This move is not without cost, however, as it creates a pressure to interpret our theories in a purely instrumentalist manner: the focus on behavioral measures and relative distrust of verbal reports makes it more difficult to interpret the

functional forms of our theories in a realist manner. If we only trust measurements of behavior, then why think that our theories capture anything other than the perception-behavior functions? Of course, this pressure does not move us entirely to behaviorism, but it does force a greater degree of explicitness about exactly which parts of our theories are supposed to correspond to actual cognitive functions.

Designing Experiments that Work

Despite the wide variety of experimental designs in psychological research, there are important, general issues about the design of controlled experiments, particularly in light of data variability and the need to (sometimes) make inferences from population-level phenomena to individual features. Suppose we want to investigate the effect of a particular treatment (e.g., the use of graphical representations) on some cognitive outcome (e.g., learning argument structure). A standard experimental approach is a between-group design with (at least) two conditions: one, the treatment, in which participants are given graphical representations, and another, the control, in which they are not. Participants are randomized to one of the conditions and the effect size is calculated from differences in some comprehension task. The randomization aims to assign participants to the treatment or control group independently of any feature of the individual or the experimental set-up. If successful, randomization ensures that any other feature that might obscure the effect of the treatment on the outcome will be equally prevalent in both the treatment and control condition, and so their effects will be balanced across the two conditions. Any observed difference between the treatment and control conditions can then be attributed to the treatment itself. (There are other advantages of randomization, such as for blinding the study, but we leave those aside here.)

This procedure is a completely standard experimental design found in many sciences. But due to the particular nature of psychological research, care needs to be taken about the inferences it supports. In particular, inferences from the experiment depend on how the effect is measured. If the effect is measured as a difference in mean between the control and the treatment condition, then the inference it supports is about a population-level phenomenon: the average effect of the treatment. No inference to any individual is necessarily licensed, since a positive difference in means can arise even if the majority of participants in the treatment group experienced a negative effect (compared to control), as long as that was outweighed by a strong positive effect of a minority. Various techniques, such as checking for extreme outliers and satisfaction of distributional assumptions of statistical tests, can help support inferences to the individual, but there is no completely satisfactory basis for inference from population-level findings to an individual. Further, even if the effect is positive for every individual in the treatment group, any inferences depend on the assumption that both the treatment and control group are representative of the wider population and – except for the treatment – of each other. This assumption will plausibly be true for large enough samples, but if the sample size is small (as it often is in psychology), we have no such assurances. It is unlikely, but not impossible, that 10 consecutive fair coin flips all come up heads. Similarly, it is unlikely-but-possible that randomization results in an assignment of treatment that is correlated with some causally relevant variable. Thus, even if we randomize, we still run the risk in small samples of not being able to identify the treatment's effects, and so we must be cautious about the correct causal assignment of the effect we observe. The possibility of spurious correlations despite randomization is not peculiar to psychology, but is exacerbated by the relatively small sample sizes.

Similarly, there are many possible explanations of a finding of no difference between the treatment and control conditions: there might not be any treatment effect, as the experimental outcome suggests, or there may be a mixture of populations in the treatment group, some for whom the treatment had a negative effect and others for which it was positive. Randomization does not help in this latter case, since if these subpopulations are suitably balanced in the general participant population, then randomization would also lead to balance across experimental conditions. In this case, the conclusion that there is no overall *average* effect is actually correct, but the inference to any particular individual might be quite incorrect. In addition, we might find no effect if the influence of the treatment is very weak relative to the noise (from any source) in the data. Randomization of treatment does not reduce outcome variance since (put intuitively) the randomization ensures only that both the treatment and the control condition contain a 'good mix' of different participants, and not necessarily a mix with low variance in performance.

In these cases, we might use an experimental design referred to as 'matching' that goes back to the philosophers Mill (1950) and Bacon (1854). The basic idea is to find two individuals that are identical on all relevant measures, but where we can observe or impose a difference in the cause variable of interest. Since the two individuals are otherwise the same, we can properly attribute any outcome differences to the treatment. More concretely, given pairs of individuals that resemble each other with respect to all variables we deem relevant, one of each pair (determined randomly) is assigned to the treatment group and the other to the control. The comparison in the outcome measurement is then performed between the matched individuals. If a significant difference is found for each pair of matched individuals (or a large proportion of them), then we have good evidence that there is an effect of the treatment on the outcome. The great advantage to this method is that the relevant variable – namely, the performance *difference*

within a pair – will typically have much smaller variance, and so we will be able to identify much smaller treatment effects.

Matching provides excellent control of the noise in the data, but there are problems for both the validity of the causal inference and inferences about the effect of treatment for an individual. The validity of the causal inference hinges on whether or not we are able to match participants properly, and whether we have matched them on all the relevant variables. If we fail to match individuals on a causally relevant variable, then we cannot attribute all effects to the treatment. Therefore, matching only works if we are in the rare situation of knowing all of the causally relevant variables for *E*, except for some uncertainty about whether the treatment is relevant for *E*. And even if the matching is perfect, differences between individuals (e.g., in strategy use) can lead to differences in outcome performance.

In practice, we can never match perfectly, and so can never completely exclude all explanations other than 'treatment causes effect' using the small sample sizes of standard psychology experiments. One might turn instead to a within-participant experimental design in which each participant is placed in every experimental condition. The match is plausibly perfect, since the same individual is compared in different conditions. Moreover, we have a large sample size because each participant can be 're-used,' and we can make claims about the treatment effect on each individual. However, the application of multiple treatments is not always possible, either for practical reasons, or because exposure to one condition changes behavior in another.

The interactions in experimental design are subtle. For noise reduction, one wants homogeneous, carefully selected participants; for causal inference from treatment to outcome, one wants a large sample with a proper randomized treatment assignment. For inference to the individual from population-level data, large sample tests require uniformity assumptions; for

tests on the individual, we need to know the other active causal effects. There is no perfect experimental design, but only a set of trade-offs involving sample size, effect size, sample noise, and prior knowledge.

Data Analysis and Null Hypothesis Testing

Even for 'perfect' data, statistical analysis faces its own problems. Often, though not always, statistical tests are used to investigate whether there is a difference in the value of a particular parameter between the treatment and control group. Results are reported and deemed publishable if a certain statistical significance (generally $p < 0.05$) is achieved. There has been a long, ongoing debate in the psychology literature on the value of Null Hypothesis Significance Tests (NHSTs) as measures of successful or relevant research (Gigerenzer, 1993; Harlow, Muliak, & Steiger, 1997; Huberty & Pike, 1999; Kline, 2004; Krantz, 1999; Oakes, 1986). The core issues arise as early as Berkson (1938), and were known to psychologists in the 1960s (Rozeboom, 1960). A ban on the use of NHSTs in psychology was discussed in the 1990s (Harlow*, et al.*, 1997; Harris, 1997; McLean & Kaufman, 1998), and the debate about alternatives to using NHSTs continues today (Harlow*, et al.*, 1997; Kline, 2004; Thompson, 1999, and many others).

In a NHST, one tests a null hypothesis, H0, of the baseline value of a particular parameter in the control population, against an alternative hypothesis, H1. H1 can either be non-specific (H1 = not-H0), or specific (H1 = parameter γ has value q). Informally, NHSTs are performed at a particular significance level α that specifies the so-called rejection region of the parameter space: if the estimate of the parameter of interest falls within this region, then H0 is rejected as a correct description of how the data was generated. The rejection region corresponds to data that is deemed so surprising in light of H0 that it is taken as sufficient to reject H0 as true. A NHST is

typically reported by a *p*-value that indicates how small the significance level α could have been such that the observed data would still have led to a rejection of H0; that is, the *p*-value is a measure of just how surprising the data would be if H0 were true. The standard threshold of $p < 0.05$ is arbitrary; there is no fundamental justification for that standard. In particular, there is no reason to think that nature's effects are all at least so strong that they can be discriminated by a test with significance level of 0.05. Moreover, with sufficient sample size, all effects become significant because the slightest approximation errors of the model become detectable.

The debate about NHSTs in psychology has mainly been concerned with the danger of misinterpretation of the results of such a statistical test. There are many other, more technical issues relating to hypothesis testing, but we will not touch on them here (though see Cox, 1958; DeGroot, 1973). At least four systematic misunderstandings of aspects of NHSTs have featured prominently in the debate:

1. The *p*-value is taken as the probability that H0 is true.
2. Rejection of H0 is taken as confirmation of H1.
3. (1-*p*) is taken as the probability that rejection of H0 will be replicated.
4. Failure to reject H0 repeatedly is interpreted as failure to replicate an earlier study.

We will briefly cover each of these errors. First, the *p*-value is technically the smallest α-value such that the observed data would result in rejection of H0. That is, a small *p*-value indicates that the data is very unlikely given H0: P(D | H0) is almost zero. But that does not imply that H0 is very unlikely given the data, since that requires that P(H0 | D) be almost zero. In general, P(H0 | D) = P(D | H0) × P(H0) / P(D), and so the *p*-value, P(D | H0), is only informative about the quantity of interest, P(H0 | D), when we also know (at least) something about the prior probability of the hypothesis, P(H0).

The second misconception is that rejection of H0 is tantamount to confirmation of H1. This fallacy arises from thinking about rejection of H0 in a NHST in the context of an over-simplified logical argument: H0 or H1; not H0; therefore H1. This argument is valid, but sound only if two conditions are met. First, H0 and H1 must exhaust the hypothesis space, which will happen only when H1 = not-H0. There might be infinitely many other alternative hypotheses that might be true or equally confirmed. For example, if H0 = '*T* and *E* are uncorrelated,' then rejection of H0 only confirms '*T* and *E* are correlated,' not any hypothesis about the specific degree of correlation. In this sense, NHSTs typically provide little positive information. Second, we must be able to test the theories in isolation from other commitments (Duhem, 1954), but essentially all NHSTs require auxiliary assumptions about the distribution of the data. Consequently, if our test assumptions (e.g. unimodality) are faulty, then we might reject H0 even though H0 is correct.

The third and fourth misinterpretations focus on the replication of experimental results. The quantity (1-*p*) does not provide any probability for replicating the observed results, largely because *p*-values depend on both the data and H0. Suppose H0 is actually true, but statistical variation led to an unrepresentative data sample from H0, which resulted in a small *p*-value. The probability of replication in this case is quite small, since an extremely unlikely event occurred, even though (1-*p*) is quite large. Alternatively, if H1 is true, then a replication probability should give the likelihood of H1 generating such a sample again. Such a probability should not depend on H0, but the *p*-value does depend on H0. More generally, it is not entirely clear what is meant by replication; surely, an exact replication of the parameter estimate (or even worse, the data) cannot be a requirement. Rather, the probability of replication should be something akin to an estimate of the likelihood of repeated rejection of H0 or H1, but (1-*p*) is not a measure of this.

Moreover, *contra* misconception #4, there are many reasons why one experiment could result in data that reject H0, while a second experiment leads to data that fails to reject H0. For example, failure to reject might be due to a sample (in either experiment) that happens to be improbable if H0 is true. More importantly, if the alternative hypotheses (H1) differ in the two experiments or if distributional features change, then the rejection region will change and consequently even the exact same data can lead to rejection in one experiment and failure of rejection in another.

These are all issues of misinterpretation and so suggest a sociological solution: simply help psychologists (as a community) have a more thorough understanding of statistical theory. However, the main issue emerging from the debate is that NHSTs do not really give the psychologist the information that she wants. In particular, psychologists are presumably most interested in: How likely is a particular hypothesis given the data, and how likely is replication of the results of a particular experiment? In fact, neither question is directly addressed by a hypothesis test as described. In more recent years, the debate on NHST has turned to providing suggestions of methods that do answer these questions (Harlow*, et al.*, 1997; Kline, 2004, and references therein). The emphasis has been on moving away from a single, numeric representation of the experimental effect and towards the use of a variety of different tools for evaluation. One suggestion has been to compute confidence intervals, which represent the set of unrejected null hypotheses and so indicate the entire space of values consistent with the data at a given significance level. However, with regard to rejection or failure of rejection, confidence intervals provide no more information than is already represented by the *p*-value, since the *p*-value corresponds to the largest confidence interval that does not include the original null hypothesis. Furthermore, confidence intervals also run the risk of serious misinterpretation: they do not provide a 1-α probability assurance that the true value of the parameter lies within the

confidence interval, but rather only assurance that the confidence interval covers the true parameter 1-α percent of the time (Belia, Fidler, Williams, & Cumming, 2005).

The other major response to these concerns about NHSTs has been a move towards Bayesian methods (Harlow*, et al.*, 1997; Kline, 2004, and references therein). Bayesian methods allow an explicit calculation of the probability of the hypothesis given the data, P(H | D), and Bayesian confidence sets provide intervals which do represent the probability of the true parameter being contained within their boundaries. However, there are concerns about using Bayesian methods. Bayesian methods require a prior probability for each hypothesis, i.e. for each hypothesis under consideration one has to specify *a priori* how likely it is. Two senses can be given to such a prior: it can either be a representation of a researcher's subjective degree of belief in each hypothesis, or it could be some kind of objective measure of the hypothesis's likelihood. In the former case, some individuals object that we should not introduce a subjective component to the science. In the latter case, it remains unclear how such objective priors can be obtained. In addition, there is some concern that these methods are much more difficult to implement, although this is becoming less of a problem.

A more general lesson from the NHST debate is that experimental results should be subject to more careful data analysis and reported more precisely, ultimately with the goal of making the claims based on experimental research more testable and/or more highly corroborated. We can clarify our results by the use of explicit models that capture the relations between multiple variables (such as structural equation models), the use of tests of model fit, and the estimation of effect sizes. Model fitting has the advantage that the aim is not to reject H0, but rather to fit the model it represents as well as possible. This avoids the misinterpretation problems that can arise from a focus on rejection. Estimation of effect sizes provides a

quantitative estimation of the difference between treatment and control conditions that can be further tested. Alternatively, various measures for corroboration of a hypothesis and measures of model fit have been suggested (McDonald, 1997), but there are currently no explicit accounts of corroboration that ensure that a highly corroborated theory is in some sense approximately true. Lastly, most scientists would agree that a result is a significant finding if it can be replicated. Consequently, rather than emphasizing the significance found in one experiment, more emphasis should be placed on the meta-analysis of several or repeated experiments. There is a vast literature on meta-analytic techniques (Kline, 2004, chapter 8, and references therein), although no particular theoretical account of replicability has found broad acceptance or usage (see Killeen, 2005 and accompanying discussions).

Conclusion

Most of the discussions and debates in the philosophy of psychology focus on high-level conceptual challenges, such as the nature of mental properties, the possibility of mental causation, and so forth. Our aim in this chapter has been to show that the more mundane, everyday aspects of the practice of psychology – experimental design, model development, statistical analysis, and so on – also provide a rich ground of conceptual and methodological challenges. The large variability of psychological data, the possibility of implicit cognitive mechanisms and individual differences, difficulty of experimental design and control, and the widespread use (and misuse) of null hypothesis statistical testing all stand to benefit from serious philosophical investigation. Almost certainly, there are no perfect solutions for any of these problems, only better and worse alternatives. But much work is needed to determine which responses are appropriate for which situations, and to find novel methods for handling these challenges.

References

Ackerman, P. L. (1988). Determinants of individual differences during skill acquisition: Cognitive abilities and information processing. *Journal of Experimental Psychology: General, 117*, 288-318.

Anderson, J. R., & Sheu, C.-F. (1995). Causal inferences as perceptual judgments. *Memory & Cognition, 23*, 510-524.

Ashby, F. G., & Gott, R. E. (1988). Decision rules in the perception and categorization of multidimensional stimuli. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 14*, 33-53.

Bacon, F. (1854). *Novum organum*: Parry & MacMillan.

Barrett, L. F., Tugade, M. M., & Engle, R. W. (2004). Individual differences in working memory capacity and dual-process theories of the mind. *Psychological Bulletin, 130*, 553-573.

Belia, S., Fidler, F., Williams, J., & Cumming, G. (2005). Researchers misunderstand confidence intervals and standard error bars. *Psychological Methods, 10*, 389-396.

Berkson, J. (1938). Some difficulties of interpretation encountered in the application of the chi-square test. *Journal of the American Statistical Association, 33*, 526-542.

Brown, S., & Heathcote, A. (2003). Bias in exponential and power function fits due to noise: Comment on Myung, Kim, and Pitt. *Memory & Cognition, 31*, 656-661.

Buehner, M. J., & Cheng, P. W. (1997). Causal induction: The Power PC theory versus the Rescorla-Wagner model. In M. G. Shafto & P. Langley (Eds.), *Proceedings of the 19th annual conference of the cognitive science society* (pp. 55-60). Mahwah, NJ: LEA Publishers.

Buehner, M. J., Cheng, P. W., & Clifford, D. (2003). From covariation to causation: A test of the assumption of causal power. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 29*, 1119-1140.

Campbell, J. (in press). An interventionist approach to causation in psychology. In A. Gopnik & L. E. Schulz (Eds.), *Causal learning: Psychology, philosophy, and computation*. Oxford: Oxford University Press.

Chi, M. T. H., Feltovich, P. J., & Glaser, R. (1981). Categorization and representation of physics problems by experts and novices. *Cognitive Science, 5*, 121-152.

Chu, T., Glymour, C., Scheines, R., & Spirtes, P. (2003). A statistical problem for inference to regulatory structure from associations of gene expression measurement with microarrays. *Bioinformatics, 19*, 1147-1152.

Colyvan, M. (in press). Naturalising normativity. In D. Braddon-Mitchell & R. Nola (Eds.), *Conceptual analysis and philosophical naturalism*. Cambridge, MA: The MIT Press.

Cox, D. R. (1958). Some problems connected with statistical inference. *The Annals of Mathematical Statistics, 29*, 357-372.

Danks, D., & Glymour, C. (2001). Linearity properties of Bayes nets with binary variables. In J. Breese & D. Koller (Eds.), *Uncertainty in artificial intelligence: Proceedings of the 17th conference* (pp. 98-104). San Francisco: Morgan Kaufmann.

DeGroot, M. H. (1973). Doing what comes naturally: Interpreting a tail area as a posterior probability or as a likelihood. *Journal of the American Statistical Association, 68*, 966-969.

Dienes, Z., Altmann, G. T. M., Kwan, L., & Goode, A. (1995). Unconscious knowledge of artificial grammars is applied strategically. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 21*, 1322-1338.

Dienes, Z., & Perner, J. (1999). A theory of implicit and explicit knowledge. *Behavioral and Brain Sciences, 22*, 735-755.

Dienes, Z., & Scott, R. (2005). Measuring unconscious knowledge: Distinguishing structural knowledge and judgment knowledge. *Psychological Research, 69*, 338-351.

Duhem, P. (1954). *La théorie physique: Son objet, sa structure* (P. P. Wiener, Trans.). Princeton, NJ: Princeton University Press.

Ericsson, K. A., & Charness, N. (1994). Expert performance: Its structure and acquisition. *American Psychologist, 49*, 725-747.

Fazio, R. H., & Olson, M. A. (2003). Implicit measures in social cognition research: Their meaning and use. *Annual Review of Psychology, 54*, 297-327.

Gigerenzer, G. (1993). The superego, the ego and the id in statistical reasoning. In G. Keren & C. Lewis (Eds.), *A handbook for data analysis in the behavioral sciences* (Vol. 1: Methodological Issues, pp. 311-339). Hillsdale, NJ: Erlbaum.

Goschke, T. (1997). Implicit learning and unconscious knowledge: Mental representation, computational mechanisms, and brain structures. In K. Lamberts & D. R. Shanks (Eds.), *Knowledge, concepts, and categories* (pp. 247-333). Hove, UK: Psychology Press.

Griffiths, T. L., & Tenenbaum, J. B. (2005). Structure and strength in causal induction. *Cognitive Psychology, 51*, 334-384.

Harlow, L. L., Muliak, S. A., & Steiger, J. H. (Eds.). (1997). *What if there were no significance tests?* Mahwah, NJ: Lawrence Erlbaum Associates.

Harré, R. (1996). AI rules: Okay? *Journal of Experimental and Theoretical Artificial Intelligence, 8*, 109-120.

Harris, R. J. (1997). Ban the significance test? *Psychological Science, 8*.

Huberty, C. J., & Pike, C. J. (1999). On some history regarding statistical testing. In B. Thompson (Ed.), *Advances in social science methodology* (Vol. 5, pp. 1-22). Stamford, CT: JAI Press.

Just, M. A., & Carpenter, P. A. (1992). A capacity theory of comprehension: Individual differences in working memory. *Psychological Review, 99*, 122-149.

Killeen, P. R. (2005). An alternative to null-hypothesis significance tests. *Psychological Science, 16*, 345-353.

Kline, R. B. (2004). *Beyond significance testing*: American Psychological Association.

Krantz, D. H. (1999). The null hypothesis testing controversy in psychology. *Journal of the American Statistical Association, 94*, 1372-1381.

Kruger, J., & Dunning, D. (1999). Unskilled and unaware of it: How difficulties in recognizing one's own incompetence lead to inflated self-assessments. *Journal of Personality and Social Psychology, 77*, 1121-1134.

Kunst-Wilson, W. R., & Zajonc, R. B. (1980). Affective discrimination of stimuli that cannot be recognized. *Science, 207*, 557-558.

Lee, M. D. (2006). A hierarchical Bayesian model of human decision-making on an optimal stopping problem. *Cognitive Science, 30*, 1-26.

Lee, M. D., & Webb, M. R. (2005). Modeling individual differences in cognition. *Psychonomic Bulletin & Review, 12*, 605-621.

Lober, K., & Shanks, D. R. (2000). Is causal induction based on causal power? Critique of Cheng (1997). *Psychological Review, 107*, 195-212.

Mandler, G., Nakamura, Y., & van Zandt, B. J. (1987). Nonspecific effects of exposure on stimuli that cannot be recognized. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 13*, 646-648.

McDonald, R. P. (1997). Goodnes of approximation in the linear model. In L. L. Harlow, S. A. Muliak & J. H. Steiger (Eds.), *What if there were no significance tests?* Mahwah, NJ: Lawrence Erlbaum Associates.

McLean, J., & Kaufman, A. S. (Eds.). (1998). *Statistical significance testing*.

Medin, D. L., & Schaffer, M. M. (1978). Context theory of classification learning. *Psychological Review, 85*, 207-238.

Mill, J. S. (1950). *Philosophy of scientific method*. New York: Hafner.

Minda, J. P., & Smith, J. D. (2001). Prototypes in category learning: The effects of category size, category structure, and stimulus complexity. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 27*, 775-799.

Minda, J. P., & Smith, J. D. (2002). Comparing prototype-based and exemplar-based accounts of category learning and attentional allocation. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 28*, 275-292.

Mitchell, D. B. (2006). Nonconscious priming after 17 years: Invulnerable implicit memory? *Psychological Science, 17*, 925-929.

Myung, I. J., Kim, C., & Pitt, M. A. (2000). Toward an explanation of the power-law artifact: Insights from response surface analysis. *Memory & Cognition, 28*, 832-840.

Navarro, D. J., Griffiths, T. L., Steyvers, M., & Lee, M. D. (2006). Modeling individual differences using Dirichlet processes. *Journal of Mathematical Psychology, 50*, 101-122.

Nisbett, R. E., & Ross, L. (1991). *The person and the situation: Perspectives of social psychology*. New York: McGraw-Hill Publishing.

Nosofsky, R. M., & Palmeri, T. J. (1998). A rule-plus-exception model for classifying objects in continuous-dimension spaces. *Psychonomic Bulletin & Review, 5*, 345-369.

Nosofsky, R. M., Palmeri, T. J., & McKinley, S. C. (1994). Rule-plus-exception model of classification learning. *Psychological Review, 101*, 53-79.

Nosofsky, R. M., & Zaki, S. R. (2002). Exemplar and prototype models revisited: Response strategies, selective attention, and stimulus generalization. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 28*, 924-940.

Oakes, M. W. (1986). *Statistical inference: A commentary for the social and behavioral sciences*. New York: Wiley.

Olsson, H., Wennerholm, P., & Lyxzèn, U. (2004). Exemplars, prototypes, and the flexibility of classification models. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 30*, 936-941.

Payne, J. W., Bettman, J. R., & Johnson, E. J. (1993). *The adaptive decision maker*. Cambridge: Cambridge University Press.

Pitt, M. A., Myung, I. J., & Zhang, S. (2002). Toward a method of selecting among computational models of cognition. *Psychological Review, 109*, 472-491.

Reber, A. S. (1967). Implicit learning of artificial grammars. *Journal of Verbal Learning and Verbal Behavior, 6*, 855-863.

Rouder, J. N., Sun, D., Speckman, P. L., Lu, J., & Zhou, D. (2003). A hierarchical Bayesian statistical framework for response time distributions. *Psychometrika, 68*, 589-606.

Rozeboom, W. W. (1960). The fallacy of the null hypothesis significance test. *Psychological Bulletin, 57*, 416-428.

Schacter, D. L. (1987). Implicit memory: History and current status. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 13*, 501-518.

Schacter, D. L., Chiu, C.-Y. P., & Ochsner, K. N. (1993). Implicit memory: A selective review. *Annual Review of Neuroscience, 16*, 159-182.

Schunn, C. D., Lovett, M. C., & Reder, L. M. (2001). Awareness and working memory in strategy adaptivity. *Memory & Cognition, 29*, 254-266.

Schunn, C. D., & Reder, L. M. (1998). Strategy adaptivity and individual differences. In D. L. Medin (Ed.), *The psychology of learning and motivation, vol. 38* (pp. 115-154). New York: Academic Press.

Shanks, D. R. (2005). Implicit learning. In K. Lamberts & R. L. Goldstone (Eds.), *Handbook of cognition* (pp. 202-220). London: Sage Publications.

Siegler, R. S. (1996). *Emerging minds: The process of change in children's thinking*. New York: Oxford University Press.

Spearman, C. (1904). The proof and measurement of association between two things. *American Journal of Psychology, 15*, 72-101.

Stanovich, K. E. (1999). *Who is rational? Studies of individual differences in reasoning*. Mahwah, NJ: Lawrence Erlbaum Associates.

Sternberg, R. J. (1977). *Intelligence, information processing, and analogical reasoning*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Taatgen, N. A. (2002). A model of individual differences in skill acquisition in the Kanfer-Ackerman air traffic control task. *Cognitive Systems Research, 3*, 103-112.

Thompson, B. (1999). If statistical significance tests are broken/misused, what practices should supplement or replace them? *Theory & Psychology, 9*, 165-181.

Tunney, R. J., & Shanks, D. R. (2003). Subjective measures of awareness and implicit cognition. *Memory & Cognition, 31*, 1060-1071.

Underwood, G. (1996). *Implicit cognition*. New York: Oxford University Press.

Whittlesea, B. W. A., & Price, J. R. (2001). Implicit/explicit memory versus analytic/nonanalytic processing: Rethinking the mere exposure effect. *Memory & Cognition, 29*, 234-246.

Wills, A. J., Reimers, S., Stewart, N., Suret, M., & McLaren, I. P. L. (2000). Tests of the ratio rule in categorization. *The Quarterly Journal of Experimental Psychology, 53A*, 983-1011.

Zaki, S. R., Nosofsky, R. M., Stanton, R. D., & Cohen, A. L. (2003). Prototype and exemplar accounts of category learning and attentional allocation: A reassessment. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 29*, 1160-1173.