Chapter 2
The Logic of Research and Questionable Research Practices: The Role of Enthymemes

Dylan R. Wong and William O'Donohue

Abstract In this chapter, we argue that poor scientific reasoning in which logical errors are made is another questionable research practice. We recommend that research psychologists and consumers of psychological research pay more attention to the logic of research by identifying the relevant inferential approaches, detecting logical errors, and constructing sound reasoning. We describe some prominent types of research logic: from alogical approaches such as that of Kuhn, to deductive logical approaches of Popper, to inductive approaches and abductive/Inference to the Best Explanation (IBE) approaches. The strength and weaknesses of each approach are discussed, along with the applications of these approaches in statistical methods and Abductive Theory of Method (ATOM).

Keywords Logic · Logical error · Questionable research practice · Clinical psychology

The Logic of Research and Questionable Research Practices: The Role of Enthymemes

Questionable research practices (QRPs) have been implicated in both creating scientific conclusions that are seen as true but are actually false (Ioannidis, 2005) and in findings that fail to replicate (see Chap. 4, this volume). One construal of QRPs is that researchers can exploit what has been called "researcher's degrees of freedom" (Simmons, Nelson, Simonsohn, 2011) that reflect choices that can shape conclusions in some desired direction. There have been a variety of QRPs identified such as selective reporting of dependent variables, *p*-hacking, hypothesizing after

D. R. Wong
Oregon Social Learning Center, Eugene, OR, USA

W. O'Donohue (✉)
Department of Psychology, University of Nevada, Reno, NV, USA
e-mail: wto@unr.edu

W. O'Donohue et al. (eds.), *Avoiding Questionable Research Practices in Applied Psychology*, https://doi.org/10.1007/978-3-031-04968-2_2

the results are known, as well as the use of the file drawer for unwanted results, and many of these are covered in this book.

Logic can be broadly defined as the study of the principles of correct reasoning and how propositions relate to one another, particularly in examining the quality of inferences from one set of propositions to another. Logic answers the question, "if Propositions P, Q, and R are true, what other propositions are also true?" Valid deductive inference has been viewed as truth preserving. Say that Propositions P, Q, and R form the premise of the argument, and Propositions Y and Z are the conclusions; if the argument is logically valid, then P, Q, and R all being true entails that Y and Z must also be true. As such logic both permits (a valid inference) and constrains (e.g., disallows fallacious inferences). Without being constrained by the limitations imposed by valid inference, researchers can infer any propositions from any other set of propositions (unlimited inferential degrees of freedom)—including making the (perhaps unwarranted) conclusion that favored views are supported. Failing to adhere to the constraints of logic may be the most fundamental QRPs, and certainly facilitates many of the other QRPs.

Scientific reasoning refers to the logical inferences made in scientific work; in the empirical sciences, for instance, researchers use some sort of reasoning to make inferences about empirical states of affairs that ought or ought not be observed given a certain theory; or to make inferences about what implications the data collected have about the truth or falsity of tested theories and hypotheses. The design of one's research can be seen as a logical exercise, that is, research design involves the construction of arguments that can entail the observational consequences of some theory, and these can then be tested to see if the propositions were valid. In addition, propositions capturing observations in their data can then be used in arguments to reason regarding whether they support or falsify other propositions. Or scholars conducting a literature review can be free to conclude what they wish. However, the actual logic of research may be obscure for psychologists: either as a normative matter (what is the best or at least a sound logic of research?) or a descriptive matter (what is the logic of this particular study?). Given that to date psychologists have paid little attention to the validity of inferences in their research, it seems fair to call any incomplete argument as *enthymemes*, a technical term meaning that the argument contains missing premises or conclusions.

Valid reasoning in sound arguments sets constraints to the researcher's degrees of freedom; it allows some conclusions to be implied and disallows many others. However, psychologists rarely, if ever, explicate the logic of their published research. In this chapter we will examine the possible choices researchers have for the logic of their research and conclude that researchers ought to be more attentive to the logic of their research and explicate their arguments better, and a failure to do so is a QRP as valid constraints imposed by logic are abrogated. We review proposals for the logic of research emanating mainly from key philosophers of science and suggest that there are several possibilities. Psychologists may have their choice on the logic of their research but should explicate these choices and be aware of their respective strengths and weaknesses.

We shall argue that standard psychological research methodologies following, say, Cook and Campbell (1979) often involve a pragmatic kind of logic. On the other hand, Popper (1959) proposed a deductive logic of research. Some other accounts of the logic of research explicitly involve inductive or abductive inferences (and its related concept of "inference to the best explanation" [IBE]), such as Haig's (2005) Abductive Theory of Method (ATOM). We shall then examine the logical inferences and errors that can also be seen in the reasoning involved in statistical methods psychologists often employ and in psychologists' pragmatic application of these, such as null hypothesis significance testing (NHST) and Bayesian inference. Finally, some accounts seem to dispense with logic altogether, such as Kuhn's (1962) account of scientific revolutions. However, each of these has limitations that must be recognized, and researchers need to be strategic in their choices for the logic of their research.

The Logic of Conventional Psychological Research

The logic of conventional psychological research might be called consistent with a weak version of pancritical rationalism (Cook & Campbell, 1979; Bartley, 1990) in that it attempts to anticipate criticisms to valid inference and promotes the design and implementation of a corresponding methodological move to potentially address that criticism. For example, let us examine the logic that is employed in the conventional double-blind randomly controlled clinical trial. Each methodological move addresses and hopefully falsifies a potential criticism/plausible rival hypothesis. For example:

1. Why the methodological move of random assignment? This at least potentially addresses the criticism that the groups differed in some systematic way *before* the experimental treatment.
2. Why the methodological move of including a no-treatment control group? This potentially addresses the criticism that due to spontaneous remission the individuals would have improved even without treatment.
3. Why double-blind? This potentially addresses the criticism that either participant expectations or experimenter expectations may have altered the values on the dependent variable(s), such that expectation effects (and not treatment effects) were responsible for such values.
4. Why a statement on the psychometric properties of the measures? This potentially addresses the criticism that the measures do not validly measure the constructs under consideration.
5. And so on, for each methodological move.

These methodological moves are supposed to be made for all "plausible rival hypotheses." But note that plausibility is a pragmatic, not a logical matter. Additionally, the question of whether the methodological move is sufficient to negate the plausible rival hypothesis also involves pragmatic judgment. Finally, it is

rare that research papers in psychology explicitly formalize the pragmatic logic that undergirds the design.

Another problem is that there are many potential criticisms/plausible rival hypotheses, and each often requires an expensive methodological move (in terms of time, subjects, and other resources). For individual studies, there may need to be a "meta-argument" regarding which the relevant priority of such criticisms—which are elements of a very large potential set of criticisms—is the most important and ought to be addressed methodologically. This may be one reason why programs of research are so important; across a series of studies more potential criticisms can eventually be addressed by including the requisite design move in at least some of the studies across the research program. For example, the randomly controlled trial described above did not address the criticism of treatment effects being caused by placebo effects; therefore, some of the subsequent studies could include an attention control condition that potentially addresses this criticism. In addition, due to the lack of follow-ups to assess for recidivism, some subsequent studies could include measurement periods of 6 or 12 months and so on. Some have commented that research sophistication grows over time as we "learn how to learn" (Munz, 2014) and an example of this may be the relatively recent concerns with clinical significance (versus statistical significance) or QRPs. This also creates a somewhat difficult problem of assessing the status of these potential criticisms across studies (e.g., studies out of Lab X did not carry out follow-up assessments but one study out of Lab Y did, however this study showed higher than desirable recidivism).

However, probably the most serious problem is that the "logic" of such research is not made clear, is not formal, and often is just inchoately pragmatic. It often plays out as an intellectual game: I would like to make a valid inference from my data to say something like "My treatment has caused improvement," and if you can present a criticism like "You can't say that because it is plausible that Z (e.g., the effects are due to placebo)," the desired causal inference is not valid. However, there are many potential criticisms and these can be leveled in an ad hoc and unsystematic way.

Deductive Reasoning

Deductive reasoning is characterized by its *demonstrability*—the use of a valid deductive inference rule establishes the truth of the conclusion if the premises are true. Thus, the conclusions of sound deductive arguments (true premises and valid deductive inference rule) are necessarily true. Another way of saying this is that valid deductive arguments are always *truth preserving*, that is, if all the premises of the argument are true and a valid deductive inference rule used, then this reasoning preserves the truth of the premises because the valid deductive inference rule always generates only true conclusions. However, a well-recognized and significant downside of deductive reasoning is that it is also *nonampliativ*e—the conclusion is not content increasing—deductive arguments simply "unpack" content that is already

contained (perhaps implicitly) in the premises of the argument. For example, consider the following deductive argument:

1. All humans are mortal.
2. Barbara is a human.
3. Therefore: Barbara is mortal.

This argument is considered nonampliative because its conclusion is implicitly contained in the first premise (because to establish that all humans are mortal one must have established that a member of this set, Barbara, is also mortal). Many early philosophers of science (e.g., Carnap, 1945) have taken deduction's nonampliative character as a sure sign that science cannot rely on deduction because science seeks new information and as such it must rely on some sort of *ampliative* reasoning—the conclusion must add or increase the information in the premises. We turn first to the view that the logic of research is deductive; this view is best exemplified by the work of Sir Karl Popper (1959).

Popperian Science

Sir Karl Popper (1959) rejected the notion coming from the logical positivists that the logic of research was inductive. Popper argued that there is no such thing as a truth-preserving ampliative inductive logic. Popper claimed that the logic of research was deductive—a *hypothetico-deductive model*—and utilized the valid logical inference rule of *modus tollens*.

In general, the logical inference rule of *modus tollens* has the following (valid/truth preserving) form:

If A, then B.
Not B.
Therefore, not A.

In science, the argument may look like the following:

1. If something is a piece of copper (A) then it conducts electricity (B).
2. This piece of copper does not conduct electricity (not B).
3. Therefore, it is not the case that all copper conducts electricity (not A).

This argument is valid because it relies on the valid logical inference rule known as *modus tollens*. To determine its soundness (i.e., the truth of its premises), the question simply becomes: are Premises 1 (the hypothesis) and 2 (the evidence) true?

Popper also suggested that formulating Premise 1 and Premise 2 ought to be guided by a few considerations: it is desirable if the conjecture being tested in Premise 1 has as great as possible *empirical content*. The empirical content of a statement is basically what it rules out. The more empirical states of affairs it rules out, the greater is a statement's empirical content. In general, scientific laws have large empirical content, ruling out many states of affairs. For example, Newton's gravitational law rules out all states of affairs except gravitational attraction

occurring in direct proportion to mass and inverse proportion to distance. As another example, "All folks in Reno, Nevada eat sugar" has less empirical content than "All Nevadans eat sugar" (i.e., empirical content increases as the number of cases it covers increases). Secondly, empirical content is increased by the precision of the statement: "All Nevadans eat at least 14 grams of sugar daily" has more precision and empirical content than "All Nevadans eat sugar."

There are also several key considerations for Premise 2, that is, the empirical test. Popper suggested it ought to be *severe*. The *severity of a test* is essentially an efficient search for the existence of falsificatory instances—cases that demonstrate the falsity of a proposition. For example, if a researcher is testing the proposition "Protestant leaders never swear," it is a more severe test to examine instances where people are most likely to swear (e.g., when they hit their thumbs with hammers, break something valuable, when someone cuts them off in traffic, etc.). It is a less severe test to examine the word use of religious leaders during sermons, or when they are teaching Sunday school, and so on, as people are generally much less likely to be disposed to swear in these situations. Thus, for Popper the research project itself should offer an argument that the test is severe. This might be such an example:

1. The most likely situations for people to swear are x, y, z.
2. People are also most likely to swear when they do not know they are being observed.
3. If my research consists of nice size samples of surreptitious sampling of x, y, z, then it is a severe test.
4. Therefore, my research project is a severe test.

Popper's overall conception of science implies that scientific knowledge can never be a matter of confirmation. Since Popperian falsification can never demonstrate the truth of theories, but can only falsify them, science progresses by eliminating its theories that are in error. Theories that survive severe testing are thought to be *corroborated* (not confirmed), for they can eventually be falsified by some future severe testing. Once these theories become eliminated, we are confronted with new problems and must build new tentative solutions subject to further falsification.

Popper's conception of science has been criticized on several grounds (O'Donohue, 2013). First, historians of science have argued that it does not reflect the historical record of how science has been practiced (Lakatos, 1970; Laudan, 1978). If Popper's goal had been to provide a description of how scientific research actually proceeds, then he has failed to do so. Second, Popper's account does not appear to address the Quine-Duhem thesis. The Quine-Duhem thesis suggests that when the falsifying event is observed (i.e., Premise 2: Not B), the initial hypothesis (Premise 1) need not be falsified; instead, the failure may be attributed to any *auxiliary hypotheses employed in the test*. Auxiliary hypotheses refer to the additional hypotheses that are required for the initial hypothesis to entail the observation. For instance, the premise "if something is a piece of copper, it conducts electricity" includes the additional premises that "the source of electricity is properly connected to the piece of copper," that "the copper is pure," and so on. According to the Quine-Duhem thesis, Popper's falsification should really take the following form:

1. If Theory and Aux_1 and Aux_2 and Aux_3… and Aux_n, then Observation.
2. Not Observation.
3. Therefore, Not (Theory and Aux_1 and Aux_2 and Aux_3… and Aux_n).
4. Therefore, Not Theory or not Aux_1 or not Aux_2 or not Aux_3 or not Aux_n.

This valid deductive argument now has an unsatisfying conclusion; along with the hypothesis being false, one could also conclude that any of the auxiliary hypotheses are also false. At least one of the hypotheses is false, but you cannot know which to blame.

Kuhn's Alogical Account of Science

Contrary to Popper, not all accounts of science claim that there is a logic of research. Kuhn's account (1962, 1994) is an example of this alogical approach, and it is noteworthy that psychologists have been particularly attentive and admiring of Kuhn's account (O'Donohue, 2013). It may even be the case that the alogical nature of Kuhn's account is partly what has drawn psychologists to his views, seeing as psychologists usually receive very little formal training in logic.

Kuhn (1962, 1994) suggested that sciences pass through several stages. In the first stage, which he called "pre-paradigmatic science," there is little progress in puzzle solving and those working in the field have deep disagreements about basic issues, for example, what constructs are important, how ought these be defined, what proper research methodology looks like, and so on. In Kuhn's second stage, someone solves a puzzle and in this puzzle solution a paradigm is born. In Kuhn's account, others in the field are impressed and influenced by this problem-solving exemplar and then begin to copy it to try to solve other problems. Scientists adopt many elements from the puzzle-solving exemplar such as its definitions, principles, methodological approaches, and so on. This becomes, for Kuhn, a "paradigm."

The field then enters a stage that Kuhn called "normal science" in which scientists attempt to apply this paradigm to solve other puzzles. According to Kuhn, sometimes these scientists are successful in puzzle solving and sometimes they are not. The problem-solving failures can accumulate and are generally frustrating to scientists. The final stage of science for Kuhn is when a scientific revolution occurs. According to Kuhn, a revolution occurs when someone applies a new approach to one or more of these failures of the old paradigm and achieves some problem-solving success. A new period of normal science then occurs where scientists ape the new paradigm until it starts the cycle all over again, that is, it produces success and anomalies and then a new revolution occurs and so on.

One can see that for Kuhn logic is not essentially involved in science. Certainly, a paradigm could in principle have a kind of reasoning; however, he is not explicit about this, and nowhere does he say that paradigms are defined by logical rules or preferences. Furthermore, new paradigms are thought to have different definitions, principles, and methodological approaches that are frequently inconsistent with the

old paradigm. Since logical rules cannot contradict one another, the kinds of reasoning in different paradigms thus cannot all be based on logical rules. Kuhn's model of research is thus alogical.

Inductive Reasoning

Given the limitations of deductive reasoning, some have looked to inductive reasoning as a good candidate for the logic of research. Induction has been taken to be an *ampliative* but *nondemonstrative* form of reasoning, that is, the conclusions of inductive arguments contain more information than their premises. However, because inductive arguments are nondemonstrative (the truth of the premises and the use of an inductive inference rule do *not* guarantee the truth of the conclusion), at best these are only *probably* true, that is, they may still be false. In the history of philosophers studying induction, their key philosophical problem has been how to quantify how likely the conclusion of an inductive argument is, given the evidence contained in the premises. Unfortunately, this problem has resisted a clear solution.

For example, notice the following about the conclusion of the following inductive argument: (1) the scope of the conclusion (helpfully) contains more information than the scope of the premises (i.e., the conclusion refers to a previously unexamined individual), and (2) even if the premises are true, the conclusion of the inductive argument might still be false—the argument is not truth preserving because no one has been able to discover a truth-preserving inductive inference rule. For example:

1. Eighty percent of the anxious subjects were successfully treated by exposure therapy.
2. Sam is anxious and will be treated with exposure therapy (Sam was not part of the anxious subjects in Premise 1, and hence was not examined to form Premise 1).
3. Therefore, Sam's anxiety will be successfully treated.

This is not a valid argument—the truth of the conclusion is not guaranteed by the truth of the premises and the inference rule used. The "problem of induction" began to concern philosophers in the nineteenth century, starting with the Scottish philosopher David Hume (1779). Hume raises the following questions: Are inferences from what is observed in the research sample to the unobserved logically justifiable? Do observed facts give us sound evidence for conclusions about similar situations that are not observed? Or, more precisely: how much evidence, if any, does the existence of an observed regularity provide toward the claim that future observations of similar phenomena will be like these past observations? The rough idea is expressed in the folk narrative that although every morning thus far the farmer has always fed the chicken, it would be false to conclude that this invariant pattern will necessarily persist—as one day the farmer will slaughter, not feed, the chicken. The future may not always be like the past.

Hume argued that *there are no nondemonstrative inferences that are also truth preserving*. Hume noted an interesting meta-paradox to his problem of induction: One cannot justify the inference deductively, because then the inference would be nonampliative. However, if one tries to justify it inductively, then it is nondemonstrative (for example, because in the past it has worked or because the probability of it working is high) and therefore one is begging the question— in other words, one is making an appeal to the very inductive principle one wishes to justify! Hume attempted to save induction by extra-logical considerations, that is, by suggesting that although induction has no logical justification, it can be based on the "natural instinct" embedded in human psychology: namely, that humans tend to expect that observed regularities will continue to occur in the future. However, this argumentative move is called "psychologism," as it is not an appeal to the logical quality of an argument, but rather it is an appeal to an alleged contingent empirical state of affairs—a hypothesized human tendency.

Hume also argues that any number of singular observations does not entail a universal statement. That is, the observation of a thousand, or even several million black crows does not entail the truth of the statement "all crows are black" because it is still at least logically possible that some yet-to-be-observed crow will turn out not to be black.

A common response to this problem has been that although no number of observations logically entails a universal statement, observations can allow a rational assignment of *some degree of (increased) probability* to the relevant conclusion. According to this view—known as *enumerative induction*—the degree of probability of the conclusion is raised upon each consistent observation. Moreover, according to this view, with many confirming instances, the inductive conclusion becomes probable to a degree that is indistinguishable or nearly indistinguishable from the certainty of a deductive conclusion.

Problems with Inductive Reasoning

Several philosophers—particularly Sir Karl Popper (1959), a notable critic of inductive reasoning and proponent of deductive reasoning in the sciences—have raised further problems with inductive reasoning.

Popper (1959) argued that the kind of observed repetition envisaged by Hume can never be perfect: the cases he has in mind cannot in principle be cases of perfect sameness; at best they can only be cases of (perhaps very high) similarity. For example, the farmer does not display the exact feeding motions each time, and there can be numerous variations in the chicken's eating. Popper argues that these are at best "repetitions" only from a certain somewhat inexact point of view. For Popper this signifies that there must always be a point of view—embodied perhaps in a system of expectations or assumptions—before there can be any perceived repetition. Popper argued:

> We must replace, for the purposes of a psychological theory of the origin of our beliefs, the naive idea of events that are similar by the idea of events to which we react by interpreting them as being similar … For even the first repetition-for-us must be based upon similarity-for-us, and therefore upon expectations--precisely the kind of thing we wished to explain. (pp. 444–445)

Popper also disagreed with the justification of induction by enumeration. If "many" consistent observations increase the probability of the universal statement, how many do we need to raise the probability to 1.0? Popper argued that universal laws (such as "All P are Q") have a large or even an infinite number of cases. Therefore, assessing the probability of a universal statement by comparing the number of tested and confirmed instances to the number of possible tests will always result in a probability of zero or near zero. Consider the proposition "All copper conducts electricity." If one estimates the number of observations of copper conducting electricity versus the number of possible observations of copper (all copper everywhere in the universe), as well as observations of observed copper but at other points in time, just because some copper once conducted electricity does not mean it always will, one can see that this fraction would essentially equal zero. Therefore, according to Popper, false theories and well-confirmed theories will have equal probabilities, that is, zero.

Induction also involves two well-known paradoxes. The first, identified by Kyburg (1961), concerns the "*lottery paradox*." Consider the following thought experiment: Suppose that there are 1000 lottery tickets numbered consecutively from one to a thousand, and that in a fair drawing one ticket has been chosen. Now let us consider the likelihood that the winning ticket is the one numbered "1." The probability that this particular ticket is the winner is only 1/1000. Therefore, the probability that some other ticket was actually drawn is 999/1000. Assuming that 0.999 is a sufficiently high probability to justify the conclusion that "some other ticket was drawn," one infers in this inductive argument that indeed some other ticket was in fact drawn. Next let us consider the ticket numbered "2." By the same reasoning we would conclude that, again, some other ticket was drawn. But notice that we can use this same reasoning for tickets numbered 3, 4, 5 … 1000. In each case, the conclusion that some other ticket was drawn seems to be confirmed by its high probability, 0.999. However—and this is where the paradox emerges—this set of conclusions is inconsistent with our knowledge that one winning ticket was actually drawn. We are thus facing a classic dilemma. Kyburg has argued that what this dilemma shows is that we cannot validly argue that something is the case simply because it has a (very) high probability of being so. Thus, there is no logic of induction.

Carl Hempel's (1965) *paradox of the ravens* points to another problem with induction. Hempel points out that the proposition "All ravens are black" is logically equivalent to the proposition, "All non-black things are nonravens." The second proposition can be logically deduced from the first using the logical law known as the law of contraposition. The law of contraposition states that "All A's are B's" is logically equivalent to "All non-B's are non-A's." Since these two propositions are logically equivalent, evidence that confirms one proposition must also confirm the

other proposition. Therefore, the observation of a white ribbon—a non-black thing that is a nonraven—would confirm the proposition that "All ravens are black." But this result is regarded as an absurdity. No one expects that a research project by an ornithologist would involve solely examining the color of, for example, ribbons. Critics of induction have taken these examples to show that certain logically proper "confirmations" seem to be substantively irrelevant.

Inference to the Best Explanation and Abduction

Inference to the Best Explanation (IBE), and its related concept of *abduction*, has been proposed as a noteworthy kind of inductive inference (Lipton, 2004). While abduction is situated in the context of *discovery* (the stage of *generating* theories and hypotheses) and IBE is situated in the context of *appraisal* (the stage of *evaluating* theories and hypotheses), both reference the same idea: namely, that one should make argumentative moves with reference to what would best explain the available evidence. With abduction, one should generate the hypotheses that have the potential to best explain the evidence, and with IBE one should evaluate the hypotheses on the basis that they best explain the evidence. We will focus primarily on IBE because the literature on IBE is far more extensive.

IBE gets its name from Gilbert Harman (1965), who defined it in the following way (p. 89):

> In making [an inference to the best explanation] one infers, from the fact that a certain hypothesis would explain the evidence, to the truth of that hypothesis. In general, there will be several hypotheses which might explain the evidence, so one must be able to reject all such alternative hypotheses before one is warranted in making the inference. Thus one infers, from the premise that a given hypothesis would provide a "better" explanation for the evidence than would any other hypothesis, to the conclusion that the given hypothesis is true.

The idea is intuitively appealing for two reasons. First, we typically believe that a fundamental aim of science is to provide explanations for phenomena. Psychologists frequently ask questions that demand explanations. Why is alcohol addictive? Why do children who experience abuse grow up to abuse their own children? Why do we think of people in social outgroups as homogenous, but see people in social ingroups as diverse? Scientific progress seems to be driven by the pursuit of explanations to such questions (Lipton, 2004), and the formation of explanations is thought to be an essential guide to the logic of research.

Second, if we consider common examples of the thoughts we have each day, we can find many examples that appear to involve IBE. Imagine hearing a voice coming from inside your house as you approach the front door. You see your housemate's car parked on the road alongside the house and notice that the door is unlocked. You infer, therefore, that the voice belongs to your housemate—perhaps they are talking on the phone. Why should you make this inference, however, and not that someone stole your housemate's car and keys, unlocked the door, and placed a speaker inside

the house playing back your housemate's voice? IBE would be the answer. This is not a deductive inference, since none of these facts *necessarily logically entail* that the voice belongs to your housemate—instead, you infer that the voice belongs to your housemate because it would be the best explanation of all the evidence available to you. It is *nondemonstrative* and *ampliative* like other inductive inferences. Lipton (2004) suggests that despite what Sherlock Holmes says he is doing in his detective work (i.e., "the art of deduction"), Holmes is actually using IBE to make his claims. He observes facts and infers to the explanation that best explains them.

Since Harman, theorists of IBE have aimed to render it in a precise analytical structure and come to a consensus regarding its validity. These efforts have been extremely challenging for two reasons. First, because IBE is an inductive kind of inference, it does not follow demonstrative rules like *modus tollens*, and as such it is unclear whether there can ever be clear rules for how IBE is to be applied. Second, it is not at all clear what is meant by *best*, and what is meant by *explanation*. While the literature on IBE has helped to develop some sense of the former, defining the latter has been extremely challenging. The history of defining explanation is quite convoluted and defining explanation remains an active topic of discussion among philosophers of science today—see Salmon's (2006) *Four Decades of Scientific Explanation* or Woodward and Ross (2021) for a comprehensive review of this history. While we often have an intuitive grasp of what it means to explain something, different explanations often have different characteristics, and developing rules for IBE that encompass all these possibilities is a tremendously ambitious project.

Despite these difficulties, several explications of IBE have been constructed: Vogel (1998), Psillos (1999), and Lipton (2004) are some prominent recent examples. The next section will briefly review some of the more prominent models of explanation, primarily drawn from Salmon (2006). After that, we describe a conception of IBE that attempts to describe what is meant by a *best* explanation through well-defined criteria and principles: Thagard's (1993) Theory of Explanatory Coherence (TEC).

Scientific Explanation

The first prominent philosopher to discuss explanation was Aristotle; he made the key distinction between *knowing-that* and *knowing-why*. The former simply involves a description of some phenomenon—"my shadow is longer in the evening"—and the latter elucidates the phenomenon—"because the sun's angle to the ground gets narrower as it sets, light travels in a straight line and my shadow is produced when my body obstructs light." This was an important first step toward understanding explanation.

However, it fell to Hempel and Oppenheim (1948; henceforth referred to as H-O)—which was further developed by Hempel (1965)—to produce first comprehensive and precise notion of explanation. According to Hempel (1965), there are

four categories of scientific explanations, shown in the table (taken from Salmon, 2006) below:

	Particular facts	General regularities
Universal scientific laws	D-N (deductive-nomological)	D-N (deductive-nomological)
Statistical scientific laws	I-S (inductive-statistical)	D-S (deductive-statistical)

The term "nomological" refers to a basic, universal scientific law. While the D-N model applies to both particular facts and general regularities, H-O only discusses the former. A D-N explanation is comprised of an *explanans* (the sentences that are to account for the phenomenon, or "the explaining sentences") and the *explanandum* (the sentence describing the phenomenon to be explained, or "the fact"). They state that an adequate D-N explanation—or a "correct answer to an explanation-seeking question" (p. 42)—must fulfill three logical conditions and one empirical condition:

1. The explanans must be a valid deductive argument (logical).
2. The explanans must contain essentially at least one general law (logical).
3. The explanans must have empirical content (logical).
4. The sentences constituting the explanation must be true (empirical).

Thus, so far we see that scientific explanation for H-O is a deductive enterprise. We have seen that Conditions 1 and 4 combined make for a valid deductive argument that leads to a true conclusion. As for Condition 2, the details of how H-O constitutes a general law are quite complicated and have been a notable weakness of the D-N model. You might notice the symmetry between the hypothetico-deductive model of Popper and the D-N model of explanation—the former uses (contradictory) evidence to falsify the law, while the latter uses the law to account for the (compatible) evidence.

One major criticism of D-N explanations as a universal model for explanations is that many satisfactory explanations do not contain scientific laws. For instance, the explanandum "I slipped on the floor" has the satisfactory explanans "the floor was wet," where both are particular facts. Though a defender of D-N explanations might suggest that the explanation is incomplete without referencing the universal scientific laws of friction, it seems like telling someone that "the floor was wet" serves as a fine explanation, and the law just serves to justify the explanation. Second, D-N explanations cannot be damaged by any number of additional premises—yet typical explanations do seem to become less useful when irrelevant premises are added to it. Third, D-N explanations are *bidirectional*, which leads to the absurdity of the explanandum explaining the explanans (e.g., my shadow being longer in the evening explains why the sun's angle is narrow). Finally, that not every fact can be explained as a scientific necessity; rather, some facts are merely probable, or statistical. Hempel (1962) attempts to address this last problem.

Statistical explanations, according to Hempel, are split into D-S and I-S. The D-S explanation is a statistical law that is deductively derived from other laws, at least

one of which is statistical. For example, explaining the outcome of a set of dice throws involves arithmetically deriving the probability of the outcome using generalizations about the dice (e.g., the probability of getting any particular die face is 1/6). The explanans of D-S explanations need not contain empirical data, and hence they are not D-N explanations. The I-S explanation follows the structure of the D-N explanation, but the law being used is statistical, and hence the explanandum is probabilistic. In Hempel's example, we might explain that someone recovered from a strep infection because they were administered penicillin, and treatment with penicillin leads to a high (e.g., 90% but not 100%) chance of recovery.

One major issue with I-S explanations is that two I-S explanations could have compatible premises but contradictory conclusions. If an individual's strep infection is resistant to penicillin, then the probability of the person recovering would be low. This would make the reference to penicillin in the explanation untenable, but the definition of the I-S explanation does not prevent this reference. Hempel attempts to correct this by, among other things, adding a further condition that the explanans must make the explanandum highly probable. But Salmon argues that the real problem lies with whether explanans changes the probability of the explanandum (e.g., whether the penicillin made the recovery from strep infection more probable). His later model of statistical relevance takes this as the fundamental condition of explanation; however, because it was later shown that causal relationships cannot be reduced to statistical relevance relationships, causal theories of explanation were later developed. Other models of explanation developed since include unificationist and pragmatic theories of explanation; the pursuit of a general model of explanation continues today (Woodward & Ross, 2021).

In response to these difficulties, some have suggested that explanation is a *primitive* concept; this means that the concept cannot be defined in terms of other concepts, and instead appeals to intuition for its characterization. There is good reason to believe this is so; after all, we can explain things ordinarily without appealing to what we mean by an explanation (see Poston, 2014, for a fuller justification for defining explanation as primitive). The Theory of Explanatory Coherence considers explanation a primitive; hence, in applying this model practically to research, we have some justification in relying on our judgment to decide whether the hypotheses we have formulated constitute an explanation.

Thagard's Theory of Explanatory Coherence

The *Theory of Explanatory Coherence* (TEC; Thagard, 1993) suggests that the "best-ness" of an explanation depends on its explanatory coherence. A theory has explanatory coherence if the propositions in the theory have explanatory relations. For instance, Propositions P and Q have explanatory coherence if one or more of the following propositions are true (p. 65):

1. P is part of the explanation of Q.
2. Q is part of the explanation of P.

3. P and Q are together part of the explanation of some R.
4. P and Q are analogous in the explanations they respectively give of some R and S.

As mentioned earlier, TEC considers explanation to be a primitive. However, Thagard also argues that TEC may be compatible with the future integration of various strands of explanation—deductive, statistical, schematic, analogical, causal, and linguistic.

TEC relies on the seven principles and three criteria. The criteria of *consilience*, *simplicity*, and *analogy* are contained within the seven principles. A theory is the most consilient if it explains the largest range of facts; he distinguishes between static consilience (the theory explains all the different types of facts) and dynamic consilience (the theory explains more types of facts than it did when it was first generated). A theory is simpler if it makes fewer "special or ad hoc assumptions" (Haig, 2005, p. 381) than other theories; this is a "check" on the consilience criterion because simpler theories tend to have lower consilience. Finally, a theory that is better supported by an analogy to previous theories is more coherent.

TEC's seven principles (Thagard, 2000, p. 43) are offered below:

1. *Symmetry.* Explanatory coherence is a symmetric relation, unlike, say, conditional probability. That is, two propositions *p* and *q* cohere with each other equally. For example:
2. *Explanation.* (a) A hypothesis coheres with what it explains, which can either be evidence or another hypothesis. (b) Hypotheses that together explain some other proposition cohere with each other. (c) The more hypotheses it takes to explain something, the lower the degree of coherence.
3. *Analogy.* Similar hypotheses that explain similar pieces of evidence cohere.
4. *Data priority.* Propositions that describe the results of observations have a degree of acceptability on their own.
5. *Contradiction.* Contradictory propositions are incoherent with each other.
6. *Competition.* If *p* and *q* both explain a proposition, and if *p* and *q* are not explanatorily connected, then *p* and *q* are incoherent with each other (*p* and *q* are explanatorily connected if one explains the other or if together they explain something).
7. *Acceptance.* The acceptability of a proposition in a system of propositions depends on its coherence with them.

As mentioned, TEC de-emphasizes prediction over explanatory coherence; instead of concerning itself with whether the theory has good predictive power *for the future* (i.e., it anticipates a set of data that is yet to be observed), it concerns itself with whether the theory has explanatory coherence *now* (based on past data and theoretical propositions). Although explanations clearly lead to predictions of certain empirical outcomes, TEC considers the latter secondary and would not abandon a theory if it led to failed predictions; this represents a contrast with Popperian science, which values predictions because they allow for possible subsequent falsification. Thagard argues that if falsification is not a good description of how the sciences actually operate (as Popperian science has been accused of; for example, see Kuhn, 1962, 1994), predictions lose value in the scientific enterprise.

A related value of explanatory coherence is that it allows one to evaluate how to modify a theory once it is falsified: if doing so would reduce the explanatory coherence of the theory (e.g., as per principle 2 of TEC, the hypothesis goes together with another hypothesis to explain another proposition), then there is stronger reason not to modify the hypothesis. Likewise, if a hypothesis contradicts a more explanatory hypothesis within the theory, then there is stronger reason to modify or remove the former hypothesis.

TEC is directly applicable to examples in psychology; for instance, Durrant and Haig (2001) apply TEC to the comparative evaluation of two theories of language. This paper compares the adaptationist hypothesis regarding language development—accordingly, humans developed language because of natural selection—and the non-adaptationist hypothesis—that it was not because of natural selection. Accordingly, they find that adaptationist accounts of language development have strong consilience (it explains many of the features of language) and simplicity (it can account for all the features of language with that hypothesis) and are supported by analogy (language resembles the development of other well-understood biological adaptations, such as the eye). Conversely, non-adaptationist hypotheses have poor explanatory coherence: they have poor consilience because they cannot account for as many features, and they have poor simplicity because they require many hypotheses to explain how each feature of language arose separately.

In short, TEC is a conception of IBE that features as a method of appraising theories. Theories with greater explanatory coherence (evaluated on the criteria and principles described above) garner greater support, and vice versa.

Applications of Scientific Reasoning

Scientific reasoning is ubiquitous in the methods we use to conduct and analyze our research. Here, we discuss one ubiquitous feature of conventional psychological research methods—Null Hypothesis Significance Testing (NHST)—and its logical flaws. We then discuss the logic of Bayesian statistics, which addresses some of the major flaws of NHST and its status as a possible alternative to NHST. Finally, we describe a recent and promising theory of science that is grounded in abductive and explanatory reasoning: Haig's Abductive Theory of Method (2005).

The Logical Flaws of Null Hypothesis Significance Testing

NHST is another ubiquitous feature of conventional psychological research, and it features centrally as a QRP. As noted in a chapter in this book by O'Connor and Khattar (Chap. 7, this volume, 2022), NHST continues to be employed by many psychologists despite its numerous and well-documented problems. In their chapter,

the authors thoroughly explore the problems associated with using NHST. In this section, we lay out NHST in the context of scientific reasoning and provide some reasons for why the use of NHST is logically flawed.

A null hypothesis is the hypothesis that the difference between two means of some variable in some population being compared is zero. NHST is thus defined in O'Connor and Khattar as follows:

> Conventionally, researchers make such decisions by assuming the null hypothesis to be true and, given this assumption, attempting to make inferences based on the probability of obtaining the actual pattern of results observed. Specifically, a statistical test yields the probability of a given results (or one more extreme) being produced by chance if the null hypothesis is true. … If this (probability) is less than a threshold probability or alpha level (typically .05), then chance is concluded to be a sufficiently unlikely explanation of the outcome, and the existence of an effect is held to be supported by the data. (Pollard & Richardson, 1987, p. 159)

NHST also relies on the following assumptions:

> 1) the null hypothesis is exactly true; 2) the sampling method is random sampling; 3) all distributional requirements, such as normality and homoscedasticity, are met; 4) the scores are independent; 5) the scores are also perfectly reliable; and 6) there is no source of error besides sampling or measurement error. (Kline, 2013, p. 74)

Homoscedasticity means that the variance in the relation between the dependent and independent variables across the different values of the independent variable is the same. For example, the relation between age and weight is often not homoscedastic, because at younger ages, the variance in weight is generally much lower than the variance in weight at adulthood.

NHST is founded on a frequentist view of probability that takes probability to be "the likelihood of an outcome over repeatable events under constant conditions except for random error" (Kline, 2013, p. 40). In other words, the probability of an event is the proportion of events occurring if the same circumstances were repeated many times (this is called the law of large numbers). This contrasts with the subjectivist view, which takes probability to be the researcher's subjective state of belief regarding the likelihood of an event—this does not rely on the event being repeatable. Frequentists consider the probability of the data given a set parameter (in NHST, this parameter is a "difference of zero"); subjectivists consider the probability of a parameter being true given that data that is set.

Conventional NHST involves both deductive and inductive reasoning. In the deductive portion, NHST assumes that the null hypothesis is true, and then deductively infers what the expected value of the test statistic should be under that assumption; the *p* value then represents the probability of obtaining the test statistic with reference to a distribution of results from simulated hypothesis studies. This is a matter of deductive logic because if the null hypothesis is true, the distribution of simulated results necessarily follows, and the *p* value follows accordingly. In the inductive portion, the researcher generalizes the comparison of the test statistic, drawing an analogy from the distribution of results of the simulated hypothesis studies to the distribution of our sample data, and drawing conclusions accordingly.

This is an inductive move because there is no guarantee that our sample data are similar to the simulated distribution; we assume this based on the fact that our sample data exhibit the six properties mentioned earlier.

As O'Connor and Khattar suggest, most problems with NHST arise from its misuse. Despite its ubiquity in psychological science, NHST is frequently misinterpreted and misapplied, and the conclusions drawn from its use are often invalid. What psychologists often hope to obtain from NHST are simply not produced by it. Some of the logical flaws in the application of NHST are listed below.

First, NHST simply tells us how likely the test statistic is likely to be obtained with reference to a distribution of imaginary test statistic values. The data from which the *p* value arises are drawn from an imaginary distribution, developed via simulation: "mathematical formulas that mimic the results from a long series of identical hypothetical studies in which the null hypothesis is true" (O'Connor & Khattar, this volume, 2022). This is problematic because the null hypothesis may not be true for our sample dataset. This disconnect leads us to make all sorts of misinterpretations regarding the *p* value: that is, that it represents "the probability of making a Type I error," or that it tells us that "5% of all published findings are Type I errors," where a Type 1 error refers to the rejection of a hypothesis when it is actually true. As O'Connor and Khattar state, if the null hypothesis is true for our sample data, then the probability of a Type I error must be zero. NHST relies on our assumption that the null hypothesis is true. Thus, the argument assumes that the null hypothesis is true, yet a conclusion is drawn about the truthfulness of the null hypothesis—this common misinterpretation is thus a logical error.

Second, psychological researchers can consistently violate the assumptions of NHST in their research: researchers may not randomly sample, sample sizes may be too small to achieve the distributional requirements of NHST (normality and homoscedasticity), and the scores obtained are never perfectly reliable (O'Connor & Khattar, this volume, 2022). For instance, Szucs and Ioannidis (2017), conducting an empirical assessment of published effect sizes and estimated power among psychology and cognitive neuroscience journals, found that the power of these studies was "unacceptably low" (p. 13). Power is defined as the probability of finding statistical significance when there is a real effect. Significance tests are constructed to produce valid results only when all the assumptions are met; as such, it is likely that the results obtained from many of the studies using NHST are biased. Viewing NHST as a form of argument, it is an invalid argument to reach the conclusion implied by the argument when the premises are untrue.

Third, the results of NHST would be insufficient for researchers to definitively attribute the difference between means to the effect that is hypothesized. The American Statistical Association statement on NHST notes that "by itself, a *p value* does not provide a good measure of evidence regarding a model or hypothesis" because it "provides limited information" (Wasserstein & Lazar, 2016, pp. 131–132). The results from NHST would still need to be combined with other background information and assumptions regarding our experiments. These assumptions may be

related to our experimental design (e.g., in a between-group experiment, the variables being controlled for did not systematically differ across groups) or our background knowledge (e.g., the background literature provides evidence that the effect being tested for in NHST is plausible). For example, consider a clinical trial investigating a novel psychotherapy for depression. If it finds that there was a significant difference between the treatment and control groups, the conclusions drawn regarding the trial still rely on other aspects of the trial: the way in which depression outcomes are operationalized, the timeframe for measuring outcomes, and so on. This idea can also be seen in the distinction between "statistically significant" and "clinically significant"—a result may be statistically significant but have little to no clinical implications. Yet the conclusions we draw frequently use the *p* value to adjudicate between the falsity and truth of the hypothesized effect in question without considering these other premises (O'Connor & Khattar, 2022). In the language of logic, it is an invalid argument to conclude that the effect is present when the truth of the premises is not clearly established.

A related key flaw of NHST is its failure to take background information regarding the effect into account. The process of NHST begins with a "blank slate" assumption—no prior information regarding the effect under consideration is considered. Each experiment and its results are considered in isolation, and the conclusion is taken as a definitive answer to the question (albeit technically subject to replication). The logical flaw underlying this problem is that prior information—such as results from previous experiments for a related hypothesis—is part of the pool of evidence from which one should infer. Using only a subset of the available evidence would likely lead researchers to a conclusion that contradicts other aspects of the pool of evidence. For example, consider that high-powered experiments have been conducted to test Hypothesis H_1, of which 1 had a positive result and 9 had a negative result (under NHST). This suggests that the prior probability of H_1 being true is 1/10. If you were to ignore this prior evidence and conduct an experiment that yields a positive result, you might wrongly conclude that H_1 is true, when it is far more likely to be a false positive (Szucs & Ioannidis, 2017). Techniques such as meta-analysis (Glass, 1976), which aggregate previous results about the effect size of a particular hypothesis to determine its robustness and value, have been developed to overcome this problem.

Finally, NHST results in an "all or nothing" outcome: the null hypothesis is either significant or not significant, and researchers often (erroneously) draw the conclusion that the effect in question is "true" or "false." By collapsing the outcome into a "clean" binary (Gelman & Carlin, 2017, p. 901), the researcher risks obscuring the uncertainty of the statistical conclusions drawn. While there is truth or falsity to a hypothesized effect, there is uncertainty inherent to every psychological experiment—some sources of uncertainty arise in experimental error, imperfect reliability and validity of measurements, and uncertainty regarding the validity of previous experiments. As such, to come to a binary conclusion regarding an experimental outcome ignores the truth that scientific methods are inherently uncertain.

Bayesian Inference

The Bayesian statistical approach, sometimes called Bayesian inference, addresses some of the flaws of NHST and has been proposed as an alternative to NHST for data analysis in psychological science. Bayesian statistics takes a subjectivist view of statistics and is founded on the mathematically precise Bayes' theorem:

$$P(H \mid E) = \frac{P(H)P(E \mid H)}{P(E)}$$

where $P(x)$ refers to the probability of x, H is the hypothesis, and E is the evidence. The notation " | " means "given"; hence, $P(H|E)$ is the probability of H given E, which is also known as the "posterior" probability of H. $P(H)$ is the prior probability of the hypothesis, $P(E|H)$ is the probability that the data are generated given that the hypothesis is true—this is sometimes called the likelihood function—and $P(E)$ is the probability of the data according to the model. $P(E)$ is also known as the "normalizing constant," which simply divides the probabilities obtained across the distribution to ensure that the distribution sums to 1. Because $P(E)$ does not figure into determining the relative probabilities of different hypotheses, the equation is sometimes depicted as:

$$P(H \mid E) \propto P(H)P(E \mid H)$$

where $\propto$ means "proportional to." By this formula, the probability of H, given the evidence, is proportional to the probability of the hypothesis multiplied by the probability of the evidence given the hypothesis. This tells us exactly how to change our degrees of belief in a hypothesis.

Once $P(H|E)$ is obtained, it "updates" the model, becoming the new prior probability for the hypothesis. Upon receiving a new set of evidence, the new prior probability replaces $P(H)$, and Bayesian updating occurs again. If $P(H|E_1)$ refers to the posterior probability of the hypothesis after receiving evidence E_1, then upon receiving new evidence E_2:

$$P(H \mid E_2) = \frac{P(H \mid E_1)P(E_2 \mid H)}{P(E_2)}$$

The subsequent equation allows the researcher to update their posterior probability of H once again. $P(H|E)$ is typically represented as a probability distribution, wherein each potential value of a given parameter implied by the hypothesis (i.e., the effect size) has a discrete probability attached to it, representing the subjective belief of the researcher in each potential value of the parameter.

The logic of Bayesian statistics is founded on the rigor of Bayes' theorem. Because Bayes' theorem holds true across any potential set of probability

distributions, it allows the researcher to determine the probability of any hypothesis so long as the probabilities corresponding to the researcher's beliefs are input into the equation. As a form of argument, it is valid because the conclusion (the posterior) will be true if the premises (the priors and the likelihood) are true.

Bayesian statistics thus holds several advantages over NHST. First, as a subjectivist view of statistics, it is a more direct way of determining the probability of a hypothesis being true given the data. Second, Bayesian estimation requires the researcher to specify their priors, which makes the researcher's preconceptions regarding the hypothesis transparent. NHST does not account for the researcher's biases, which can and do influence the results obtained (see Chap. 5 on *p*-hacking in this volume). Third, the product of Bayesian estimation is a probability distribution of parameter values representing the degrees of belief in the effect being tested; it does not commit to a binary yes/no outcome. This thus allows for uncertainty to be represented. Fourth, Bayes' theorem explicitly incorporates prior information in the estimation by requiring that prior probabilities are introduced as inputs. The information from previous experiments regarding a hypothesized effect can thus be accounted for. Fifth, it provides a way of precisely updating the probabilities upon receiving new evidence, allowing the researcher to determine exactly how their degrees of belief should change considering the new evidence. Meta-analyses can estimate the effect size of a hypothesis, but they often rely on hundreds of studies to do so. Finally, the fundamental logic of Bayesian statistics implies that the posterior distribution found will probably change as new evidence is introduced through further research; the researcher cannot help but be reminded that the results they obtain are "pending." Results from NHST have often been presented and understood as the definitive answer regarding a hypothesis, despite the rhetorical emphasis on replication.

Here is a simple example of Bayesian inference, borrowed from Kruschke (2015). Suppose that you are trying to find the bias of a coin. Based on prior information, for example, this coin came from a magician's shop, and the shopkeeper tells you that the coin mostly lands on heads, you suspect that the coin is strongly biased toward heads. As such, you might hypothesize that the coin's bias is 0.9, where 0 represents tails and 1 represents head. However, because you are not completely confident about this, you construct a prior distribution wherein the prior probability distribution is densest in the region around 0.9—in this example, a distribution known as the beta distribution is the most appropriate. This probability distribution represents your prior knowledge and your confidence in your hypothesis, $P(H)$. Subsequently, you conduct some tests—you flip the coin ten times, finding that it lands on heads eight times. What should you believe about the bias of the coin?

Bayesian inference, through Bayes' theorem, allows you to determine this precisely. To find $P(E|H)$, you ask: what is the probability of obtaining eight heads in ten coin flips if we suspect the true bias of the coin is 0.9? For example, calculating the discrete probability for the point estimate of 0.9:

$$P(E \mid H) = 0.9^8 \times 0.1^2$$
$$= 4.3 \times 10^{-3}$$

The equation above represents the fact that the event with a probability of 0.9 (heads) occurs eight times and the event with the probability of 0.1 (tails) occurs twice. You would then ask the same question for every discrete point estimate in the prior probability distribution, obtaining $P(E|H)$ for each value. Finally, using Bayes' theorem, you can input the values of $P(H)$ and $P(E|H)$ into the proportional relationship, allowing you to find the precise probabilities for each point estimate in the distribution. Doing these calculations by hand is computationally intensive, but many statistical programs now have implementations of Bayesian statistics that are quite efficient at applying it.

One major criticism of Bayesian statistics is that the choice of the prior distribution is arbitrary. Since the prior distribution strongly influences the effect of evidence on the posterior distribution, the posterior distribution obtained may be heavily (and incorrectly) biased. For instance, assume that a coin is strongly biased toward heads, that is, the coin has a true parameter value of 0.9, where 0 represents tails and 1 represents head. If I have a strong belief that the coin has no bias, for example, centers the prior probability distribution narrowly around the parameter value of 0.5, then the appearance of nine heads among ten coin flips will still lead to a posterior distribution that clusters near the initial prior (e.g., 60% biased toward heads). This strong belief is referred to as a "strongly informative prior."

However, in practice, researchers using Bayesian estimation would choose a prior based on the background information available to them. For instance, if they had no information regarding the coin, they might choose a flat prior, making all possible parameter values have equal probability. Nine heads in ten coin flips would then lead to a posterior distribution centered around the true parameter value. Alternatively, being aware that the coin belonged to a magician whose trick relied on the coin turning up heads, I might set a prior probability distribution that clusters around a parameter value of 1. This is also close to the true parameter value of the coin. Additionally, the prior distribution becomes more likely to converge on the true value of the parameter over time, suggesting that there exists a sufficient number of observations for the likelihood function to overwhelm even a strongly informative prior.

There is some controversy as to the status of Bayesian inference as a deductive or inductive method (see Gelman, 2011; Talbott, 2008). Bayes' theorem is clearly deductive since it relies on the deductive rules of mathematics. It follows the logic of *modus ponens*: namely, that so long as the premises are true, the conclusion is true. The truth of the conclusion is contained in the truth of the premises. So long as the premises are true, the deductions obtained from Bayes' theorem are valid.

Hawthorne (1993) argues that Bayesian inference is a form of *eliminative induction*, or "induction by deduction." As evidence that is deductively entailed by the

hypothesis builds up, some hypotheses get eliminated (probability reduced to zero) and one (the true hypothesis) rises to the top as all other alternatives are eliminated. In cases where evidence is only probabilistically related to the hypotheses, some hypotheses get "highly refuted" and one (the true hypothesis) becomes "highly confirmed" as its plausibility increases. In both cases, Bayesian inference is thought to result in convergence to agreement regarding the posterior probabilities of hypotheses. The former may appear to be Popperian falsification, which implies that a potentially infinite number of hypotheses to be falsified prevents us from ever knowing the true hypothesis. In response, Hawthorne suggests that if hypotheses are "ordered" in plausibility, so long as the hypotheses above the true hypothesis in the order are "evidentially distinguishable" (evidence exists that can deductively show that one hypothesis is true and another false), the true hypothesis will eventually rise to the top of the order and remain there.

IBE in Haig's Abductive Theory of Method

Brian D. Haig is a cognitive psychologist and research methodologist who advocates the use of both abduction and IBE in behavioral sciences. These feature in his *Abductive Theory of Method* (ATOM; Haig, 2005), whereby abduction is the primary tool for generating theories and IBE is the primary tool for appraising these. Haig's theory is comprehensive and detailed, and space limitations prevent a complete discussion of it here—as such, a quick sketch of ATOM will be laid out, focusing on the place of IBE within it. ATOM uses TEC for its grounding, committing to its notions of explanation as a primitive and the distinctions between explanation and prediction.

ATOM centers on the principle that explanatory considerations play a role across the three stages of theory construction: theory generation, theory development, and theory appraisal. Theory development occurs after theory generation, and theory appraisal occurs throughout the generated theory's lifespan. Unlike hypothetico-deductive models of science such as Popper's (1959)—that begins with a problem and a theory aimed at solving this problem—ATSM begins with the phenomena to be explained and suggests that theories are constructed based on the phenomena. Accordingly, "phenomena exist to be explained rather than serve as the objects of prediction in theory testing" (Haig, 2005, p. 371). Importantly, Haig distinguishes phenomena from data—data are the raw observations, while phenomena are the "robust empirical regularities" (p. 372) that are abstracted from the data—in ATSM, they are also called *phenomenal laws*. Haig provides some examples of phenomenal laws from psychology: "the matching law, the Flynn effect in intergenerational gains in IQ, and the recency effect in human memory" (p. 374).

In the theory generation stage, abductive inference—reasoning to underlying causal mechanisms to explain phenomena—is used to judge the plausibility of potential causal mechanisms, and the best of these are then selected as a "plausible

enough" theory. This judgment of plausibility is based on its explanatory value, which is evaluated using the criteria of explanatory coherence. Here, *existential abduction* is applied, wherein the existence of previously unknown objects and constructs is hypothesized. This is contrasted with *analogical abduction*, whereby models of the mechanisms are developed based on analogy to other known mechanisms; this is used in the theory development stage. For instance, if it is theorized that anatomical and physiological patterns in different generations of animals can be explained by the theory of natural selection, an analogical model would be that of artificial selection.

Finally, theory appraisal involves the use of IBE. In contrast to the Popperian model of the logic of research, that evaluates the theory based on its survival from falsification attempts, the ATOM model judges the theory based on its "explanatory breadth" (p. 380), which is synonymous with Thagard's (1993) criterion of consilience. Furthermore, unlike the Bayesian model of confirmation that relies on assigning probability to various hypotheses in light of evidence, the ATOM model judges on Thagard's (1993) qualitative explanatory criteria, not quantitative statistical criteria—note that this contrasts with the justification of IBE based on simulations previously explored.

ATOM is thought to be a particularly useful philosophical contribution for clinical psychologists for three reasons. First, it was developed for application to the behavioral sciences; Popper, Kuhn, and other prominent theorists of science based their models on the physical sciences, especially physics (O'Donohue, 2013), and as such their models may not be applicable to the behavioral sciences. Second, it is a theory founded in the practice of science; it pays attention to all of the steps involved in scientific activity (theory generation, theory development, theory appraisal)—again, Popper and others have been accused of not basing their models of science on the actual practice of scientists. Finally, and quite intriguingly, Ward et al. (2016) have elaborated that ATOM can be integrated into the practices of clinical psychologists as a conceptual framework for psychological assessment.

One criticism of the theory appraisal stage of ATOM (Romeijn, 2008) is that it is subject to two common objections to IBE, labeled by Lipton (2004) as "Hungerford's objection" and "Voltaire's objection." Hungerford's objection suggests that the notion of "best-ness" of explanations is too subjective and varied. However, given the grounding of ATOM in Thagard's (1993) IBE, which has been naturalistically justified (subject to empirical testing), Romejin is willing to concede this point. Voltaire's objection suggests that there is no reason to believe that the theories chosen by IBE are true or approximately true, for we have no reason to believe that the world accords with our explanatory criteria. To that point, Haig (2008) responds that ATOM does not claim to be a method for revealing truths; instead, the explanatory criteria in TEC are guides to truth, or at least would bring us toward the goal of "maximizing true propositions and minimizing false ones" (p. 1042).

Conclusions

We argue that poor scientific reasoning in which logical errors are made is another questionable research practice. We recommend that research psychologists and consumers of psychological research pay more attention to the logic of research by identifying the relevant inferential approaches, detecting logical errors, and constructing sound reasoning. We describe some prominent types of research logic: from alogical approaches such as that of Kuhn, to deductive logical approaches of Popper, to inductive approaches and abductive/IBE approaches. The strength and weaknesses of each approach are discussed, along with the applications of these approaches in statistical methods and ATOM.

References

Bartley, W. W. (1990). The retreat to commitment. Open Court.

Carnap, R. (1945). On inductive logic. *Philosophy of Science, 12*(2), 72–97. https://doi.org/10.1086/286851

Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation: Design & analysis issues for field settings*. Houghton Mifflin.

Durrant, R., & Haig, B. D. (2001). How to pursue the adaptationist program in psychology. *Philosophical Psychology, 14*(4), 357–380. https://doi.org/10.1080/09515080120088067

Gelman, A. (2011). Induction and deduction in Bayesian data analysis. *Rationality, Markets and Morals, 2*, 67–78, 1999.

Gelman, A., & Carlin, J. (2017). Some natural solutions to the *p* -value communication problem – And why they won't work. *Journal of the American Statistical Association, 112*(519), 899–901. https://doi.org/10.1080/01621459.2017.1311263

Glass, G. V. (1976). Primary, secondary, and meta-analysis of research. *Educational Researcher, 5*(10), 3–8. https://doi.org/10.3102/0013189X005010003

Haig, B. D. (2005). An abductive theory of scientific method. *Psychological Methods, 10*(4), 371–388. https://doi.org/10.1037/1082-989X.10.4.371

Haig, B. D. (2008). On the permissiveness of the abductive theory of method. *Journal of Clinical Psychology, 64*(9), 1037–1045. https://doi.org/10.1002/jclp.20507

Harman, G. (1965). The inference to the best explanation. *The Philosophical Review, 74*, 88–95.

Hawthorne, J. (1993). Bayesian induction is eliminative induction. *Philosophical Topics, 21*(1), 99–138. https://doi.org/10.5840/philtopics19932117

Hempel, C. G. (1965). *Aspects of scientific explanation*. Free Press.

Hume, D. (1779). An enquiry concerning human understanding. In D. Hume (Ed.), *Essays and treatises on several subjects, Vol. 2. Containing an enquiry concerning human understanding, a dissertation on the passions, an enquiry concerning the principles of morals, and the natural history of religion* (pp. 3–212). Unknown Publisher. https://doi.org/10.1037/11713-001

Hempel, C. G. (1962). Deductive-nomological vs. statistical explanation. University of Minnesota Press, Minneapolis. Retrieved from the University of Minnesota Digital Conservancy, https://hdl.handle.net/11299/184632

Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLoS Medicine, 2*(8), e124. https://doi.org/10.1371/journal.pmed.0020124

Kline, R. B. (2013). *Beyond significance testing: Statistics reform in the behavioral sciences* (2nd ed.). American Psychological Association. https://doi.org/10.1037/14136-000

Kruschke, J. K. (2015). *Doing Bayesian data analysis: A tutorial with R, JAGS, and Stan (Edition 2)*. Academic Press.
Kuhn, T. S. (1962). *The structure of scientific revolutions*. University of Chicago Press.
Kuhn, T. S. (1994). *The structure of scientific revolutions* (2nd ed.). University of Chicago Press.
Kyburg, H. E. (1961). *Probability and the logic of rational belief*. Wesleyan University Press.
Lakatos, I. (1970). Falsification and the methodology of scientific research programmes. In I. Lakatos & A. Musgrave (Eds.), *Criticism and the growth of knowledge*. Cambridge University Press.
Laudan, L. (1978). *Progress and its problems: Towards a theory of scientific growth* (1st ed.). University of California Press.
Lipton, P. (2004). *Inference to the best explanation* (2nd ed.). Routledge/Taylor and Francis Group.
Munz, P. (2014). *Our knowledge of the growth of knowledge: Popper or Wittgenstein?* Routledge.
O'Connor, B. P. O., & Khattar, N. (2022). Controversies regarding null hypothesis significance testing. In M. Lillienfeld & O'Donohue (Eds.), *Questionable research practice: Designing, conducting, and reporting sound research in clinical psychology*. Springer.
O'Donohue, W. (2013). Clinical psychology and the philosophy of science. *Springer International Publishing*. https://doi.org/10.1007/978-3-319-00185-2
Pollard, P., & Richardson, J. T. E. (1987). On the probability of making Type I errors. *Psychological Bulletin, 102*, 159–163. https://doi.org/10.1037/0033-2909.102.1.159
Popper, K. R. (1959). *The logic of scientific discovery*. Hutchinson.
Poston, T. (2014). *Reason and explanation*. Palgrave Macmillan. https://doi.org/10.1057/9781137012265
Psillos, S. (1999). *Scientific realism: How science tracks truth*. Routledge.
Romeijn, J.-W. (2008). The all-too-flexible abductive method: ATOM's normative status. *Journal of Clinical Psychology, 64*(9), 1023–1036. https://doi.org/10.1002/jclp.20516
Salmon, W. C. (2006). *Four decades of scientific explanation* (1st ed.). University of Pittsburgh Press.
Szucs, D., & Ioannidis, J. P. A. (2017). Empirical assessment of published effect sizes and power in the recent cognitive neuroscience and psychology literature. *PLOS Biology, 15*(3), e2000797. https://doi.org/10.1371/journal.pbio.2000797
Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science, 22*(11), 1359–1366. https://doi.org/10.1177/0956797611417632
Talbott, W. (2008). Bayesian epistemology. In Zalta, E. N. (Ed.), *The Stanford encyclopedia of philosophy* (Winter 2016 Edition). https://plato.stanford.edu/archives/win2016/entries/epistemology-bayesian
Thagard, P. (1993). *Conceptual revolutions* (1st ed.). Princeton Univ. Press.
Thagard, P. (2000). *Coherence in thought and action*. MIT Press.
Vogel, J. (1998). Inference to the best explanation. In E. Craig (Ed.), *Routledge encyclopedia of philosophy*. Routledge. https://www.rep.routledge.com/articles/inferenceto-the-best-explanation
Ward, T., Clack, S., & Haig, B. D. (2016). The abductive theory of method: Scientific inquiry and clinical practice. *Behaviour Change, 33*(4), 212–231. https://doi.org/10.1017/bec.2017.1
Wasserstein, R. L., & Lazar, N. A. (2016). The ASA statement on p -values: Context, process, and purpose. *The American Statistician, 70*(2), 129–133. https://doi.org/10.1080/00031305.2016.1154108
Woodward, J., & Ross, L. (2021). Scientific explanation. In Stanford Encyclopedia of Philosophy. https://plato.stanford.edu/archives/sum2021/entries/scientific-explanation