Black Box Inference: When Should Intervening Variables Be Postulated?

Elliott Sober

ABSTRACT

An empirical procedure is suggested for testing a model that postulates variables that intervene between observed causes and observed effects against a model that includes no such postulate. The procedure is applied to two experiments in psychology. One involves a conditioning regimen that leads to response generalization; the other concerns the question of whether chimpanzees have a theory of mind.

1 Introduction
You observe a number of causes impinge on a system. You also observe the system generate a number of effects. When should you infer that the causes are linked directly to the effects, and when should you infer that causes are linked to effects by passing through intervening variables?

Perhaps the most familiar instance of this problem is the debate in psychology between mentalism and radical behaviourism. Mentalists hold that beliefs, desires, and sensations are caused by environmental contingencies and themselves cause behaviour. Radical behaviourists claim that there is no need to postulate such intervening variables—that a model linking stimulus
directly to response is preferable. However salient this example might be, the problem of black box inference is more general. The ‘system’ on which causes impinge and that generates effects may be an organism, but it also may be bigger or smaller. Economies experience inputs and outputs; so do organs and cells. The problem of black box inference isn’t just about organisms and their behaviour; it also arises at higher and lower levels of organization.

Just as black box inference occurs in domains outside of psychology, so it is not restricted to the question of whether an intervening variable should be postulated. To be sure, it is natural to begin with the question of choosing between 0 such postulates and 1. However, it would not be surprising if similar issues arise when the question is 2 versus 1. The question of whether to postulate an intervening variable is a special case of the question of how many such variables a theory should invoke. This means that the problem of black box inference should interest the convinced cognitivist. Even if it is a mistake to forswear intervening variables unconditionally, it still is worth asking when and why intervening variables should be introduced.

2 One cause and one effect
Let us begin, then, with the following simple problem. You observe one cause (C) and one effect (E), and your question is which of the models in Figure 1 is better.

![Figure 1](https://example.com/figure1.png)

In describing radical behaviourists as declining to postulate intervening variables, I mainly have in mind the position of B.F. Skinner [1938, 1950]. It should be noted, however, that many behaviourists had no problem with the idea of intervening variables and that Skinner himself occasionally said that he had nothing against them.

E. C. Tolman introduced the term ‘intervening variable’ into psychology. Many psychologists now have the impression, due largely to the influential article by MacCorquodale and Meehl [1948], that Tolman thought that intervening variables must be operationally defined in terms of observables. However, Tolman [1948] explicitly rejected that interpretation; see Amundson [1983] and Smith [1986] for further discussion. Clark Hull [1943] also held that intervening variables are legitimate theoretical entities, although he saw no reason to interpret them as cognitive representations.

The intervening variables postulated by the models I’ll discuss in what follows are not operationally defined, in the sense that their existence and states cannot be deduced from input and output data. Rather, the framework is probabilistic; intervening variable models make probabilistic predictions that differ from those made by models that do not postulate intervening variables. This probabilistic approach was anything but alien to Tolman; see, for example, Tolman and Brunswik [1935] and Brunswik [1943]. Brunswik was influenced by Reichenbach (as am I); Tolman and Brunswik discovered that their approaches had a great deal in common (Smith [1986]).
I call the first model ‘N-l, 1’ because it postulates no intervening variable and describes the relation of one cause and one effect. The second model postulates an intervening variable that connects one cause to one effect.

You might be tempted to opt for the intervening variable model by appealing to the principle of *no action at a distance*. If the events C and E are temporally separated, then the causal connection of C to E must be mediated by a causal process in which there are intermediate links (Hull [1943]; Hempel [1965], pp. 203–4). To this argument, it may be replied that the (N-l, 1) model is committed to action at a distance no more than (IV-1, 1, 1) is. The latter model neglects to assert that the connection of C to I and the connection of I to E is mediated by further events, but that is not the same as denying that such events exist. The distinction between agnosticism and atheism needs to be borne in mind.

Although I think that this reply is correct, it has the consequence that the two models are not incompatible; rather, the relationship is that (IV-1, 1, 1) entails (N-l, 1), but not conversely. This raises the question of whether they are, properly speaking, competitors, a question that I will not pursue further. I will note, however, that if they do not compete with each other, then there can be no reason (other than pragmatic reasons of convenience) to accept the one and reject the other.

One factor that does separate the two models is parsimony—the (N-l, 1) model postulates fewer events and causal connections than the (IV-1, 1, 1) model. In conventional scientific practice, this fact about the two models is taken to confer on the (N-l, 1) model the status of a *null hypothesis*—it is presumed innocent until proved guilty. If the two models fit the observations about equally well, you should opt for (N-l, 1); you should abandon the simpler model in favour of the more complex alternative only if the (IV-1, 1, 1) model does a better job of accommodating the observations.

The question, then, is whether the intervening-variable model fits the observations better. Let’s begin by characterizing the observations. In black box inference, we observe frequencies—f(C), f(not-C), f(E|C), f(E|not-C), f(E), and f(not-E). From these, we must infer values for the probabilities postulated by the models. The usual procedure is *maximum likelihood estimation*. The best estimate of a set of probabilities is the one that maximizes the probability of the observations. For example, if E occurs 72% of the time when C occurs in your observational sample, then the best estimate of Pr(E|C) is 0.72.

---

2 If an arrow in a causal model entailed that there must be an intervening variable, this would have the consequence that causal chains must be dense rather than quantized. I see no reason to think that causal models in different sciences automatically have commitments on this fairly recondite matter.

3 Since maximum likelihood estimation ignores prior probabilities, Bayesians do not think much of it. However, the points I’ll make in what follows still stand within a Bayesian framework, so long as you distinguish what the current data tell you about the competing models from the prior information you have about them.
The thing to notice about maximum likelihood estimation in the case of the 
(N-1, 1) model is that there is a unique maximum-likelihood assignment of 
values. For each observed frequency, there is a probability in the model whose 
value needs to be fixed, and there is one such assignment that does the job best, 
namely the one in which the probability is assigned a value that matches the 
observed frequency.

The (IV-1, 1, 1) model is different. For example, if E occurs 72% of the time 
when C occurs in a set of observations, the maximum likelihood estimate of 
Pr(E|C) is 0.72. However, according to the intervening-variable model, that 
probability is a function of three other independent quantities:

$$Pr(E|C) = Pr(E|I)Pr(I|C) + Pr(E|not-I)Pr(not-I|C)$$

There now are too many unknowns. As a result, the (IV-1, 1, 1) model fails to 
be identifiable.4

What is the significance of this fact? If you insist that a model must 
be identifiable, then this is a reason to favour the (N-1, 1) model over the 
(IV-1, 1, 1) model. If you do not insist on this, you nonetheless might want to 
compare the likeliest version of the (N-1, 1) model with any of the likeliest 
versions of the (IV-1, 1, 1) model, and ascertain which of them is likelier—i.e. 
which confers a higher probability on the observations. The answer is that the 
models tie. As we’ve just seen, if the maximum likelihood estimate of Pr(E|C) 
in the first model sets that parameter equal to 0.72, the second model can do 
just as well, but no better, by choosing values for Pr(E|I), Pr(E|not-I), and 
Pr(I|C), so that Pr(E|C) comes out having a value of 0.72.

In summary, the (N-1, 1) model has the advantage of being simpler and 
also of being identifiable. From the point of view of accommodating the data, 
the two models do equally well. We have identified no ground on which the 
(IV-1, 1, 1) model is better than the (N-1, 1) model.5

3 Multiple causes, multiple effects, and multiple intervening 
variables

The models just considered treat C, I, and E as dichotomous variables, but 
the result can be generalized. Suppose there are three dichotomous causal

4 To put this point more carefully, the (IV-1, 1, 1) model has five independently adjustable 
parameters (Pr(C), Pr(I|C), Pr(I|not-C), Pr(E|I), and Pr(E|not-I)), whereas frequency information 
about C and E furnishes only three independent observations; you observe f(E), f(C), f(E|C), and 
f(E|not-C), but the first of these frequencies is a function of the remaining three. This is why (IV- 
1, 1, 1) is not identifiable.

5 Of course, one might have background knowledge that leads one to think that the intervening-
variable model is better in this instance. However, the focus of the present paper is on how the 
data and features intrinsic to the models considered permit a choice to be made. This is why I so 
far have restricted my attention to how parsimonious the models are and how probable they say 
the observations are.
variables ($C_1, C_2, C_3$) and three dichotomous effects ($E_1, E_2, E_3$). Consider two models. The first (Figure 2) postulates no intervening variable; rather, it connects each cause with each effect. The second model (Figure 3) says that each effect is influenced by a single intervening variable that itself has the three causes as inputs.

As before, the model without the intervening variables is simpler and identifiable; in terms of accommodating the data, the two models do equally well.

### 4 Multiple causes, multiple effects, and a single intervening variable

How, then, is it possible for intervening variable theories to be superior to theories that refuse to postulate these hypothetical entities? Whiten [1993, 1994, 1995], building on ideas he found in Miller [1959] and Hinde [1970], suggests that an intervening variable theory will be preferable to a theory that posits a direct connection between causes and effects when the intervening variable model postulates a single intervening variable mediating the connection of multiple causes to multiple effects (Figure 4).
Whiten points out that this model has six 'causal linkages'; in contrast, the (N-3, 3) model, which connects each of three causes directly to each of three effects, has nine. He says that the intervening-variable model is preferable because it is 'more economic of representational resources' ([1995], p. 284).

To illustrate this general point, Whiten discusses the question of why we are entitled to attribute the intervening variable of 'thirst' to an organism. Postulating this internal state has the advantage of simplifying one's account of how different stimuli are related to different responses (Figure 5). By invoking 'thirst' as an intervening variable, 'the multitude of S–R links no longer needs to be coded. The observer needs only to know that the "Ss" can each lead to the state (thirst, in this case) and that for an organism in this state, certain "Rs" can be forecasted' (Whiten [1995], p. 284).

Although Whiten counts 'causal linkages', rather than parameters, to determine how parsimonious these two models are, the bottom line for this example is the same if all variables are dichotomous. In this case, the intervening variable model has eight parameters for probabilities of the form \( \Pr(I|\pm C_1 \& \pm C_2 \& \pm C_3) \); for each \( E_i \), there will be one parameter for \( \Pr(E_i|I) \) and another for \( \Pr(E_i|\text{not-}I) \). The (IV-3, 1, 3) model thus has \( 8 + 2 + 2 + 2 = 14 \) adjustable parameters. The (N-3, 3) model contains 24 parameters.

One reason to count parameters in a model, not arrows in a diagram, is that the former procedure is invariant over different graphical representations, whereas the second is not. The (N-3, 3) model is diagrammed above with 9 arrows; however, the model can be represented equivalently in terms of a single 8-state causal variable (Figure 7). Now the no-intervening variable model has three 'causal linkages'. However, when the (IV-3, 1, 3) model is represented in terms of the variable C 8, it has 4 causal links (Figure 7).
The star representations make the intervening variable model seem less parsimonious; the unstarred representations convey the opposite impression. To see which impression should be trusted, we need to consider the underlying probability model that these diagrams represent. The intervening-variable model has fewer adjustable parameters.

The intervening-variable model would not be more parsimonious if the I variable had the same number of states as all the causes together. For example, suppose that each of the three C’s and each of the three E’s is dichotomous, but that I can occupy eight states. Then, for each of the eight I states, there will be eight parameters of the form Pr(I_i | ±C_1 & ±C_2 & ±C_3). And for each E_j, there will be 8 parameters of the form Pr(E_j | I_i). The total, then, for this intervening-variable model is (8)(8) + (8)(3) = 88. If the right way to evaluate parsimony is to count adjustable parameters in a probability model, not the number of arrows in a diagram, then the intervening-variable model is more economical only to the extent that the intervening variable represents an informational bottleneck; there must be fewer I states than C states, and fewer I states than E states. This entails that there must be a loss of information in the transition from input causes to the intervening variable. The procedure of counting

The idea of an informational bottleneck has another sort of significance. Bamber and Van Santen [forthcoming] consider models in which causal arrows do not entail probability increases, and define a notion of testability according to which a model is testable if and only if there exists at least one data set that would refute it. They show that a model with a single intervening variable is testable in that sense precisely when the intervening variable represents an informational bottleneck. They do not consider the comparative question of when an intervening-variable model will be better supported than a no-intervening-variable model.
arrows in a diagram is not sensitive to this point, which is another reason why model evaluation should be based more directly on the probability model itself.

Whiten says that the intervening-variable model is better on grounds of parsimony, but doesn’t say why parsimony matters in this particular inference problem. Presumably, he is relying on the general idea that parsimony is always a virtue of scientific theories, and so requires no special defence in this instance. Even if we suppose that this is so, a second issue needs to be addressed. Notice that the (IV-3, 1, 3) model is always more parsimonious than the (N-3, 3) model, assuming that I involves an informational bottleneck. Does this mean that an intervening-variable model is always preferable, regardless of what the data happen to be? As already noted, there is more to consider in model selection than how parsimonious the competing theories are. A quite separate consideration is how well the competing models fit the data at hand. Even if simplicity is a tie-breaker when two theories fit the data equally well, it still can turn out that the more complex model is preferable because it fits the data significantly better than its simpler alternative. It also can turn out that the simpler model fits the data better than its more complex rival, so that appeal to simplicity becomes unnecessary in justifying the choice of the simpler alternative.

5 Conditional independence

There is a predictive difference between the (IV-3, 1, 3) and the (N-3, 3) models. The no-intervening-variable model predicts that the different effects should be conditionally independent of each other—that is, that \( \Pr(E_i \& E_j | \pm C_1 \& \pm C_2 \& \pm C_3) = \Pr(E_i | \pm C_1 \& \pm C_2 \& \pm C_3) \Pr(E_j | \pm C_1 \& \pm C_2 \& \pm C_3) \) (for all \( i \neq j \)). The intervening-variable model, on the other hand, predicts that the observed effects should fail to be conditionally independent of each other. This prediction follows on quite general and standard assumptions, one of which is that the system is not deterministic—probabilities must be strictly between 0 and 1.

The Akaike information criterion explains how fit-to-data and parsimony (as measured by the number of adjustable parameters in a model) jointly affect a model’s predictive accuracy; see Akaike [1973] and Forster and Sober [1994] for discussion. When two models are nested, the simpler model can’t fit the data better. For example, consider a model that says that \( y \) is a linear function of \( x \) and one that says that \( y \) is a parabolic function of \( x \):

\[
\begin{align*}
(LIN) & \quad y = a + bx \\
(PAR) & \quad y = a + bx + cx^2.
\end{align*}
\]

Since (LIN) is a special case of (PAR), obtained by setting \( c = 0 \), (PAR) must always fit the data at least as well as (LIN) does. However, in the case of (IV-3, 1, 3) and (N-3, 3), the models are not nested, which means that it is perfectly possible that the former should fit some data sets better.

If \( C \) causes \( I \), and \( I \), in turn, causes both \( E_1 \) and \( E_2 \), then \( E_1 \) and \( E_2 \) will not be independent of each other, conditional on \( C \), if (i) \( I \) screens off \( E_1 \) from \( E_2 \), (ii) \( I \) screens off \( C \) from each \( E_i \), (iii) all probabilities are intermediate, (iv) \( \Pr(E_i | I) \neq \Pr(E_i | \text{not-}I) \). The proof is given in Appendix A of Sober [1994b].
Whether conditional independence obtains can be answered by gathering frequency data on the C’s and E’s. If \(E_1\) occurs 1/5 of the time when \(C_1 \& C_2 \& C_3\) occurs, and \(E_2\) occurs 1/4 of the time when \(C_1 \& C_2 \& C_3\) occurs, does the conjunction \(E_1 \& E_2\) occur 1/20 of the time when \(C_1 \& C_2 \& C_3\) occurs? If \(f(E_1 \& E_2)\) in this circumstance differs significantly from the product of \(f(E_1)\) and \(f(E_2)\), one should reject the hypothesis that the two effects are conditionally independent.

A slight modification of a humorous example of Reichenbach’s [1956] illustrates the logic of this argument. A theatre company travels around the country; the actors often eat together in greasy spoons. Because of this, the actors in the troupe often come down with gastro-intestinal distress—say, once every hundred days. This is true of the leading man and lady in the troupe, but with this added fact—when one of them gets sick, the other almost always does too. The explanation of why their sick days are correlated is that they not only eat in the same restaurants, which is true of the entire troupe, but, being lovebirds, they usually order the same food.

Let’s suppose that the restaurants that the troupe frequents always have some tainted food on their menus; one of the hundred items on the menu is spoiled. An actor drawing at random from the menu has a 1/100 chance of eating tainted food. If two actors order their food independently, the probability that both will get sick on a given day is \((1/100)(1/100)\). This idea of independent draws may be represented by the model in Figure 8.

Here C describes the restaurant menu from which each actor draws at random. An alternative model, one that predicts that the sick days will be correlated, postulates an intervening variable (Figure 9). According to this
model, the two actors usually agree (I) as to which item on the menu they both
will order. The intervening variable model (IV-1, 1, 2) predicts that E1 and E2
will not be independent of each other, conditional on C; the no-intervening
variable model (N-1, 2) makes the opposite prediction.

This predictive difference between an intervening variable model and a
no-intervening variable model obtains, no matter how many causes there are,
as long as there are at least two effects. What, then, of the case in which there
are multiple causes and just one effect? There is no expectation in causal
modelling that a joint effect will render its different causes probabilistically
independent of each other. And as we saw before, there is nothing much to
the difference between three dichotomous causes and one eight-state cause.
The case of many causes and one effect is just like the case of one cause and
one effect, and we were hard pressed in that instance to see what advantage an
intervening variable model might possess.

Does this mean that the individuation of causes is a matter of convention,
whereas the individuation of effects is not, as far as the test of conditional
independence is concerned? In fact, just as three dichotomous causes can be
redescribed as a single eight-state cause, so the idea of two dichotomous effects
is equivalent to that of a single four-state effect. For example, according to the
(N-1, 2) model, Pr(E1 & E2|C) = Pr(E1|C)Pr(E2|C). This says that E1 and E2
are independent of each other, conditional on C, regardless of whether we
choose to say that ±E1 and ±E2 are distinct events or that ±E1 & ±E2 is a
single event.

6 Multiple layers of intervening variables

It is obvious how the framework just sketched allows more than one interven-
ing variable to be postulated. Perhaps I1 is legitimately introduced as a variable
that mediates the connection of C1 and the effects E1 and E2 while I2 similarly
turns out to be required to explain how C2 is connected to the effects E3 and
E4. However, it may be less than transparent how the test of probabilistic

10 Furthermore, the conditions required for an effect to render its causes conditionally independent
are very stringent, if the frequencies of causes change with time; see Sober [1994c] for
discussion.

11 An argument made by Skinner ([1953], p. 35) is worth considering in the light of the
methodology I have presented. Skinner considers a causal chain in which the first and third
links are, respectively, an observable cause and an observable effect. Here is what Skinner says
about the idea of postulating the existence of a second and unobservable link between them:
'Unless there is a weak spot in our causal chain so that the second link is not lawfully determined
by the first, or the third by the second, then the first and third links must be lawfully related. If we
must always go back beyond the second link for prediction and control, we may avoid many
tiresome and exhausting digressions by examining the third link as a function of the first'
(quoted in Hempel [1965], p. 186). Skinner's comment about the debilitating character of
certain intellectual activities may be set to one side; the important point to see is that Skinner's
argument is right for the case of one cause and one effect, but does not generalize.
independence can license the postulation of multiple layers of intervening variables for a single set of observed causes and observed effects.

The lesson we extracted from considering the case of one cause and one effect applies to the question of multiple layers. Suppose we observe that E1 and E2 are not independent of each other, conditional on C. This leads us to prefer (TV-1, 1, 2) over (N-1, 2). However, there is no point to multiplying intervening variables beyond necessity, which is what the model in Figure 10 does. This two-layer model can't do better than (TV-1, 1, 2), for the same reason that (TV-1, 1, 1) can't do better than (N-1, 1). The intervening variable II in this two-layer model connects one cause (C) to one effect (I2) and there is nothing to be gained from introducing a postulate of this sort.

However, if we follow the maxim that an intervening variable must always have at least two effects, it is easy to show how the data may lend support to a model that has more than one layer of intervening variable. Consider, for example, the difference between the two models in Figure 11. These models
both predict that E1 and E2 will not be probabilistically independent of each other, conditional on C (and similarly for E3 and E4). However, they disagree about how E1 and E3 should be related (and also on the relationship of E2 and E4); the (IV−1, 2, 4) model predicts independence while the (IV−1, 1, 2, 4) model makes the opposite prediction. The same observational methodology that indicates whether we should postulate one intervening variable rather than none also can say whether we should postulate two layers of intervening variables rather than one.

7 Background assumptions in black box inference

The various models I have discussed are ‘directed graphs’—they describe variables that come in two or more states and claim that causal relations obtain among them. I have suggested that these models can be tested by looking at frequency data. For this to be so, the causal claims made by the model must be taken to imply claims about probabilities. How are causality and probability connected?

Causal modelling in the sciences forges a link between these two concepts by way of two background assumptions. Causal modellers talk of the Markov condition (Davis [1985]; Spirtes et al. [1993]); philosophers will see in this idea the views that Reichenbach [1956] developed in his discussion of what he termed the principle of the common cause (Sober [1994b]). Let me divide this idea into two parts. Suppose that C occurs at time t and that E1 and E2 occur sometime later. Then:

If C is the only cause at t of E1 and of E2, and if C is the most recent common cause of E1 and E2, then Pr(E1 & E2|C) = Pr(E1|C)Pr(E2|C).

Similarly, suppose that C occurs at t1, that I occurs at t2, and that E occurs at t3. Then:

If C is a cause of I, and I is the only cause (at time t2) of E, then Pr(E|I&C) = Pr(E|I).

The first idea is that the most recent common cause of two events renders them conditionally probabilistically independent of each other, when that common cause is completely specified. The common cause ‘screens off’ the one from the other; the probabilistic equality can be formulated equivalently as Pr(E1|C & E2) = Pr(E1|C). The second principle says that a completely specified proximate cause of an effect screens off more distal causes from that effect.

I have formulated these two claims about causality by talking about complete descriptions of causal facts. Without this rider, the two principles
would be false. If C1 is a common cause of E1 and E2, C1 will fail to screen off E1 from E2 when C2 is also a common cause of E1 and E2 (Sober [1994b]; see Figure 12). And if C influences E by its pathway through I1 and by its pathway through I2, then I1 will not screen off C from E (Figure 13).

If causal claims have probabilistic implications only when they have this requisite sort of "completeness", this constrains how causal models must be interpreted if they are to be testable in the light of observed frequencies. A naïve suggestion would be to interpret the arrows in a causal model as merely attempting to describe some of the causes that the specified effects have. For example, when scientists suggest that smoking is a common cause of lung cancer and of emphysema, this does not mean that smoking must be the only common cause. The interpretative point is well taken as it applies to the verbal formulation, but it does not do justice to what causal models actually claim. The common causes postulated in a causal model are to be understood as completely specified, and the same holds for the proximate causes that a model postulates as intervening between distal causes and effects.

I have claimed that the (N-U 2) model predicts that E1 and E2 will be probabilistically independent of each other, conditional on C. And I have claimed that the (IV-1, 1, 2) model makes precisely the opposite prediction. These interpretative points do not derive just from the two causal principles stated above; in addition, one needs the idea that causal models should be interpreted as providing complete specifications of common causes and of

---

12 Even with the rider of completeness, the principles do not have the status of a priori principles; although they are standard fare across the sciences, the Bell inequality in quantum mechanics shows that they may in fact be false in that domain; see van Fraassen [1982] for an elementary exposition.
proximate causes. This may seem to impose an impossibly high standard on causal modelling, but in fact the standard is no higher than it should be. If a model fails in the respects just canvassed, then it should be replaced by something more complete.

8 Learning

Although my discussion of the issue of when intervening variables should be postulated has proceeded at a rather lofty level of abstraction, I think this formalism applies at a relatively more concrete level of psychological theorizing. In particular, it bears on how one should understand the phenomenon of learning.

Consider the relatively simple example of a pigeon in a Skinner box that is conditioned to peck when a light goes on. At the start of the experiment the bird's probability of pecking (E) is the same whether or not the light is on (C): \( \Pr(E|C) = \Pr(E|\neg C) \). At the end of the process \( \Pr(E|C) \) is large while \( \Pr(E|\neg C) \) is small, owing to the occurrence of positive reinforcement. A radical behaviourist might hold that this change in the organism's behaviour can be explained without postulating a variable that intervenes between stimulus and response. And we saw at the start of this paper that neither fitting the data nor considerations of parsimony favour an intervening-variable model, when the task is to connect one cause and one effect.

Matters change when we consider the phenomena of stimulus generalization and response generalization. In the former, an organism who is conditioned to perform \( E \) when \( C_1 \) is present simultaneously acquires the disposition to perform \( E \) when \( C_2 \) is present; in the latter, an organism who is conditioned to perform \( E_1 \) when \( C \) is present simultaneously acquires the disposition to perform \( E_2 \) when \( C \) is present. Stimulus generalization requires us to consider how multiple causes of an effect should be modelled; response generalization requires us to contemplate how multiple effects of a cause should be modelled. And we saw earlier that the logic of black box inference changes when we move from the case of one cause and one effect to cases in which there are multiple causes and multiple effects.

9 Response generalization

An interesting example of response generalization is provided by an experiment performed by Wickens [1938]; it is described in Gallistel [1980] and Enç [1995]. Subjects had their hands strapped, palm down, to a board, with the middle finger resting on an electrode. When a buzzer was sounded at the start of the experiment (with no electric shock occurring), 7 of the 18 subjects lifted
their fingers. Subjects then experienced a buzzer sound followed by an electric shock, which caused them to lift their fingers. They differed with respect to how many repetitions of buzzer followed by shock it took for them to acquire the disposition to raise their fingers when they heard a buzzer that was not accompanied by electric shock, but 12 out of 18 did so with fewer than 200 repetitions and the longest run it took anyone to acquire the disposition was 311. As soon as this disposition was in place, the subjects were placed in a second experiment, which was just like the first, except that their hands were strapped palm up. When the second experiment began and the buzzer sounded with no electric shock, 10 out of 18 subjects raised their fingers; after 20 repetitions of buzzer followed by shock, 16 out of 18 acquired the disposition to raise their fingers when they heard the buzzer but did not receive a shock.

What is interesting about these two experiments is that the muscular movements conditioned in the first experiment are the opposite of those elicited in the second. With palm down, raising your finger away from the electrode involves exciting the extensor muscles and inhibiting the flexor muscles (call this combination of muscular movements M1). With palm up, raising your finger away from the board means exciting the flexor muscles and inhibiting the extensor muscles (call this M2).

Wickens ([1938], p. 108) notes 'a very definite tendency for the conditioned response to be carried over from the one situation to the opposite.' Not only did a number of subjects lift their finger at the sound of the buzzer that began the second experiment; in addition, most subjects exhibited the disposition to do this sooner in the second experiment than they did in the first. It therefore seems natural to regard this study as an example of response generalization. It also seems reasonable to suspect that the right model for this pair of experiments involves an intervening variable. Conditioning in the first experiment modifies some internal state, which influences behaviour at the end of the first experiment and also during the second experiment.

Let us now try to move from these intuitive first approximations to something more precise. Unfortunately, Wickens' study did not treat the subjects in the two experiments to the same conditioning process; in both experiments, subjects varied with respect to how many repetitions of buzzer-plus-shock they experienced. In order to simplify my analysis, I'm going to suppose that all subjects experienced 275 repetitions in the first experiment and that all then experienced 20 repetitions in the second experiment. Wickens notes that subjects differed with respect to the strength of the conditioned response they developed, but, again for simplicity, I'm going to interpret his data as recording whether or not subjects produced a conditioned response at various points in the experiments.
Given these simplifying assumptions, we may summarize Wickens’ data as depicted in Figure 14.

As indicated, 16/18 is (statistically) significantly larger than 7/18, which allows us to conclude that the conditioning experienced during the first experiment changed subjects’ dispositions to produce M1.

Did conditioning during the first experiment also affect subjects’ dispositions to produce M2 at the start of the second experiment? That is, does the experiment furnish evidence of immediate transfer? To answer this question, we need to compare the frequency with which subjects produced M2 at the start of the second experiment after going through the first experiment (10/18) with the frequency with which they would have produced M2 if they had not gone through the first experiment. Unfortunately, Wickens did not use a control group to determine the value of this latter quantity. However, it is not unreasonable to suppose that the first and the second experiments are entirely symmetrical, and so the data from the first experiment furnishes the appropriate base rate for the second. Since 10/18 is not significantly greater than 7/18 (nor is it significantly greater than 9/18, which may be thought, a priori, to be the appropriate base rate), we cannot reject the null hypothesis that there was no immediate transfer from the first experiment to the start of the second.

As Wickens notes, subjects got conditioned faster in the second experiment than they did in the first; 16 of the 18 subjects took only 20 repetitions in the second experiment to produce M2, while it took 275 repetitions for 16 out of 18 to produce M1 in the first experiment. This is evidence for the existence of delayed transfer from the first experiment to the second, but let us explore more carefully the logic on which this conclusion is based. As just noted, Wickens did not use a control group to determine how many subjects would have raised their fingers at various points during the second experiment if they had not gone through the first experiment. However, since 7 of the 18 subjects...
in the first experiment produced M1 after 20 repetitions, we may infer that 7 out of 18 would have produced M2 in the second experiment after 20 repetitions, if they had not gone through the first experiment. Since 16/18 is significantly greater than 7/18, the conclusion may be drawn that going through the first experiment caused subjects to learn faster in the second.

Given that going through the first experiment modified subjects' dispositions to perform M1 at the end of that experiment and also influenced their performance of M2 during the second experiment, it may seem that this is evidence for the existence of an intervening variable. However, the existence of learning during the first experiment and of transfer into the second is not enough for one to favour (IV-1, 1, 2) over (N-1, 2); the issue of independence must be considered.

Let's begin by considering the data concerning performance at the end of the first experiment and at the start of the second:

\[
f(\text{subjects who produced M1 at the end of the first experiment and M2 at the start of the second}) = \frac{9}{18} = 0.5.
\]

\[
f(\text{subjects who produced M1 at the end of first experiment}) f(\text{subjects who produced M2 at the start of the second experiment}) = \frac{16}{18} \cdot \frac{10}{18} = 0.49.
\]

Since 0.5 is not significantly greater than 0.49, we cannot reject the null hypothesis that performing M1 at the end of the first experiment and performing M2 at the start of the second are independent. Not surprisingly, there is no evidence for the existence of an intervening variable in this instance.

Let us now consider the data that pertains to delayed transfer:

\[
f(\text{subjects who produced M1 at the end of the first experiment and M2 at the end of the second}) = \frac{15}{18} = 0.83
\]

\[
f(\text{subjects who produced M1 at the end of first experiment}) f(\text{subjects who produced M2 at the end of the second experiment}) = \frac{16}{18} \cdot \frac{16}{18} = 0.79
\]

Given that there were 18 subjects in the experiment, this is not a significant difference, so once again, one cannot reject the null hypothesis of independence. The data do not support the claim that there is an intervening variable in this case either.\(^{13}\)

We may summarize this analysis as follows: there is evidence for delayed transfer.

\(^{13}\) A complete test of the independence hypothesis would have to consider all possible settings of the causal variables. Let 'nC1' denote n repetitions in the first experiment and 'mC2' denote m repetitions in the second. For a given choice of values for n and m, one would like to determine whether M1 and M2 are independent of each other, conditional on each of the following: nC1 followed by 0C2, mC2 followed by nC1, mC2 followed by 0C1, and mC2 followed by nC1. Since Wickens ran his experiment only in one direction (C1 followed by C2), there are no data for two of these treatment combinations; however, symmetry considerations may make it plausible to assume that the data that would have been obtained for mC2 followed by nC1 would have been the same as the data that were in fact obtained for mC1 followed by nC2.
transfer in this experiment, but no evidence for the existence of an intervening variable. This may sound paradoxical; the reader may be inclined to insist that if the treatment in the first experiment (C) influenced what subjects did at the end of the first experiment (E1) and also what they did at the end of the second (E2), surely there must exist an internal state (I) that intervenes between the one cause and the two effects. The point is that this natural-sounding conclusion is not inevitable, and that the data do not license it in this instance.

In order to dispel the impression of paradox that may hover over this analysis, it may help to consider the following analogy. Suppose a doctor gives 18 patients a series of allergy injections. Before the series of shots is begun, 11 individuals test positive for allergies to dust and to ragweed; the other 7 test negative for both. After the treatment, it is found that 2 of the 18 test positive for dust allergy and 2 test positive for ragweed allergy. If the tests are reasonably reliable, one may conclude that the shots were effective with respect to both conditions. The question then arises whether the biochemical pathway by which the shots improved dust tolerance and the pathway by which they improved tolerance of ragweed overlapped. Is there a common intervening variable in the two processes, or did the processes proceed independently? Of the individuals who test positive for at least one allergy at the end of the experiment, suppose that one tests positive for both, one tests positive for ragweed allergy but not for dust, and one tests positive for dust allergy but not for ragweed. These data do not permit one to reject the null hypothesis that the allergy shots produced their two effects independently. (N−1, 2) is a better model of what happened than (IV−1, 1, 2).

In Wickens’ study, what sort of data set would allow one to reject the null hypothesis of independence, and so conclude that an intervening variable exists? If all subjects produce M1 at the end of the first experiment and all produce M2 at some benchmark point in the second experiment, then the null hypothesis cannot be rejected. However, if 50% of the subjects achieve criterion by the end of the first experiment, 50% do so at a chosen point in the second, and these are the same individuals, then one has evidence for nonindependence. An experiment in which all subjects behave the same does not provide evidence for an intervening variable.14

As noted earlier, the attractiveness of an intervening-variable model depends on its connecting causes to multiple effects; if there is just one effect, there is no reason to prefer the intervening-variable model (or at least I have not been able to identify such a reason). In the light of this point, the following question may be formulated:15 why should one describe Wickens’

---

14 The same point holds if the study records the frequencies of M1 during some trial period at the end of the first experiment and the frequencies of M2 during some trial period during the second experiment; if all subjects exhibit the same frequencies, the null hypothesis cannot be rejected.

15 I’m grateful to Robert Pennock for raising this point.
two-part experiment as involving two muscular movements, M1 and M2? Instead, why not say that there is a single action—moving the middle finger away from the electrode—that the subjects performed in both experiments? This redescription has the result that the two models look like those in Figure 15. Now there appears to be no reason to favour the intervening variable model.

The mere fact that there is a physical difference between M1 and M2 is not a reason to treat them as distinct effects. And the fact that subjects may be unaware that their muscles are behaving differently in the two cases is likewise no reason to treat them the same. What is crucial here is that M1 and M2 have different probabilities of occurring in the circumstances of the experiment itself. M1 becomes quite probable and M2 quite improbable by the end of the first experiment; yet, M2 is more probable than M1 at the start of the second. These differences are reflected in observable frequencies, and these frequencies should be recognized in the underlying probability model. Thus, the appropriate comparison should be between \((N-1, 2)\) and \((IV-1, 1, 2)\), not between \((N^*)\) and \((IV^*)\).

The situation here may be quite different from what occurs in many conditioning experiments. For example, suppose a pigeon is conditioned to peck when a light goes on. Let’s suppose that a pigeon can use many different muscular movements to peck at a button. In the experiment, the probability of pecking when the light is on changes; however, the experiment may not provide any evidence that the organism discriminates between the different muscular movements \((M1, M2, \ldots)\) that can underwrite that behaviour. It may be that \(Pr(M1|\text{the pigeon pecks}) = Pr(M2|\text{the pigeon pecks})\), and that neither of these probabilities changes value in the course of the experiment. If so, it would be a mistake to describe this experiment as involving response generalization.\(^\text{16}\)

Gallistel [1980] and Enç [1995] have argued that one obtains different predictions about what will happen in the second experiment, depending on

\[^{16}\text{The same point applies to the case of stimulus generalization. Suppose that pigeons are colour-blind and that a pigeon, once conditioned to peck when a red light is on, also is disposed to peck when a green light is on. This is not a case of stimulus generalization in any meaningful sense.}\]
how one describes what the subjects learned in the first. If the subjects were conditioned to move their fingers away from the electrode in the first experiment, then the sound of the buzzer at the start of the second should lead them to move their fingers away from the electrode. On the other hand, if the subjects were conditioned to excite their extensor muscles and inhibit their flexor muscles in the first experiment, then the sound of the buzzer at the start of the second should lead them not to move their finger away from the electrode. The outcome of the experiment reveals that the former hypothesis is correct; Gallistel and Enç take this to illustrate a quite general and important fact about how behaviour is organized. I do not dispute this point, but it should not be taken to mean that the right model of the experiment is one that refuses to describe the physical difference between the first behaviour and the second. That impoverished description would not distinguish between the interesting experiment that Wickens ran and the boring experiment of simply running the first half of Wickens' experiment twice. The response generalization that occurred in Wickens' experiment is interesting because it involves a generalization over different muscular movements.

10 Chimpanzee theory of mind

Although the phenomenon of response generalization is a natural setting in which to apply the test of probabilistic independence to decide whether an intervening variable should be introduced, it might seem at first glance that the phenomenon of stimulus generalization is much less amenable to this treatment. After all, in stimulus generalization, an organism who learns to produce $E$ when $C_1$ is present simultaneously acquires the inclination to produce $E$ when $C_2$ occurs. If, indeed, stimulus generalization is limited to cases in which there are two (or more) causes but only one effect, it won't be possible to use the test of probabilistic independence to discriminate between an intervening-variable model and a no-intervening-variable model. The reason is that the criterion of independence applies only when there is more than one effect term. However, as we now will see, stimulus generalization does involve multiple effects, and so the criterion of probabilistic independence applies to it just as much as it applies to the case of response generalization.

Whiten [1993, 1994, 1995] discussed the general issue of why one should accept an intervening-variable model because of its specific relevance to the question of whether nonhuman primates can be said to have a 'theory of mind'. An organism is said to have a theory of mind if it can form beliefs and desires about the mental states of other organisms (or about its own mental states). In defining the theory-of-mind hypothesis in this way, I set to one side
the issue of how the organism comes to formulate such representations—whether by 'simulation' or by some other process. If an organism has a theory of mind, it has second-order intentionality; if the organism has beliefs about objects and states of affairs in the world, but has no beliefs about mental states, the organism is said to have first-order intentionality only (Dennett [1983]).

Povinelli et al. [1990] designed one of the most interesting experiments that addresses the question of whether chimps have a theory of mind. A chimp looks through a window at two trainers in a room. One trainer (the 'guesser') leaves, while the other (the 'knower') places food into one of four containers (the chimp can't see which). The guesser then re-enters the room. The knower points to the container where the food is hidden, the guesser points to a different container, and the chimp has to choose between the knower and the guesser. The chimp then gets to search the container singled out by the trainer she chooses, and to keep the contents. After a number of rounds, three of the four chimps in the experiment learned to point to the knower, not to the guesser. The hypothesis under test is that these chimps represented the knowledge states of the two trainers.

The chimps then went into a new experiment. Now the knower and the guesser both remain in the room while the food is hidden by a third party; the knower watches the food being hidden, while the guesser wears a bag over his head. The chimps in this experiment initially acted as if they were learning from scratch—their success rates were merely at the level of chance (Povinelli [1994])—but they subsequently learned to discriminate knower from guesser in this new set-up, so that their mean success rate in the second experiment reached approximately the same level they had attained in the first experiment, and did so with less training than was required in the first experiment. Heyes [1994, forthcoming] and Povinelli [1994] conclude that the failure of immediate transfer from the first to the second experiment means that these experiments fail to provide evidence that the chimps attributed mental states to the trainers. Evidently, the chimps learned faster in the second experiment than in the first and this is evidence that they learned something in the first experiment, which they carried forward; however, it is possible to explain this delayed transfer without attributing a theory of mind to the chimps.

Let's see what these experiments look like when we ask them to discriminate between an intervening-variable model and a model that postulates no intervening variable. First some notation:

E1 = pointing to the trainer who hid the food in the first experiment
E2 = pointing to the trainer who watched the food being hidden in the second experiment
The frequencies that Povinelli et al. obtained are shown in Figure 16.

Although Povinelli et al. ran these experiments using only four chimps, the sample size was in fact much larger. The frequency of E1 in the first experiment was measured in blocks of 50 trials per chimp. The frequencies of E2 in the second experiment were measured in blocks of 10 trials per chimp. With these sample sizes, one can conclude that the average success rate at the end of the first experiment was significantly greater than the baseline with which the chimps began; the first experiment modified the chimps' probability of producing E1.

Did going through the first experiment influence the chimps' behaviour at the start of the second? That is, was there immediate transfer? To answer this question, we'd have to compare the frequency with which the chimps produced E2 at the start of the second experiment after going through the first (0.575) with how frequently they'd have produced E2 if they had not gone through the first experiment. Povinelli et al. did not use a control group to determine the value of this quantity, but symmetry considerations may suggest that an appropriate estimate is furnished by the frequency with which the chimps produced E1 at the start of the first experiment (or perhaps we should assume, a priori that the appropriate baseline is 0.5). If so, we may conclude that there is no evidence for immediate transfer, since 0.575 is not significantly greater than 0.53, nor is 0.575 significantly greater than 0.5.

Finally, an argument can be made for the existence of delayed transfer; the chimps reached a 65% success rate in only 30 trials in the second experiment, whereas it took them much longer to do that well in the first. This argument

---

Although the sample size is much larger than four when one counts each behaviour of each chimp as a single observation, one can't escape the fact that there are just four chimps in this experiment. A large number of trials may allow one to reject a null hypothesis concerning these four chimps, but will not do so if the null hypothesis concerns some species-wide characteristic.
depends on the same assumption about base rates that we discussed in connection with Wickens' [1938] study. To see whether the chimps learned faster in the second experiment because of the experiences they had in the first, one needs to compare how well they did at the end of the second experiment with how well they would have done if they had not gone through the first experiment. This requires a control group that was not part of the experimental design; however, it may be reasonable to regard the performance of the chimps at the start of the first experiment as providing a reasonable estimate. If so, we have evidence for delayed transfer, but not for immediate transfer.

Do causes render effects conditionally independent? This question can't be answered by looking just at average frequencies across chimps, but requires examination of frequencies for each chimp separately. Figure 17 depicts the frequencies with which different chimpanzees chose the knower (some provided by Povinelli, personal communication).

From these data we can compute the following:

\[
f(E_1 \text{ at the end of the first experiment and } E_2 \text{ at the start of the second}) = \frac{[(0.74)(0.6) + (0.67)(0.6) + (0.6)(0.6) + (0.67)(0.5)]}{4} = 0.38.
\]

\[
f(E_1 \text{ at the end of the first experiment}) f(E_2 \text{ at the start of the second experiment}) = (0.64)(0.575) = 0.37.
\]

Since 0.38 is not significantly greater than 0.37, we cannot reject the null hypothesis that performing \(E_1\) at the end of the first experiment and performing \(E_2\) at the start of the second are independent. Not surprisingly, there is no evidence for the existence of an intervening variable in this instance.

Let us now consider the data that pertain to delayed transfer:

\[
f(E_1 \text{ at the end of the first experiment and } E_2 \text{ at the end of the second}) = \frac{[(0.74)(0.9) + (0.67)(0.6) + (0.6)(0.8) + (0.67)(0.3)]}{4} = 0.44.
\]

\[
f(E_1 \text{ at the end of the first experiment}) f(E_2 \text{ at the end of the second experiment}) = (0.64)(0.65) = 0.42.
\]
Once again, the difference isn’t statistically significant. Even though there is evidence that delayed transfer occurs, we have no evidence that this transfer was mediated by an intervening variable.

Heyes [1994, forthcoming] and Povinelli [1994] agree that the question of whether chimps have a theory of mind is to be addressed by seeing whether there is immediate transfer or only delayed transfer; if there had been immediate transfer, this would have been evidence that chimps have a theory of mind, but since the transfer was delayed, the evidence should not be so interpreted. However, from the present point of view, this criterion seems to be off the mark. What is relevant, I suggest, is not when transfer occurs, but whether transfer is accompanied by a failure of conditional independence.

It is true that delayed transfer in this experiment can be explained without appeal to the theory-of-mind hypothesis, in the sense that the data do not rule out alternative hypotheses. However, the same can be said of immediate transfer; if it had occurred, this too would have been consistent with alternatives to the theory-of-mind hypothesis. In any event, it is hard to see why the plausibility of the theory-of-mind hypothesis depends on when transfer occurs; having a theory of mind and catching on right away are not so tightly connected. On the other hand, if the chimps are using a theory of mind in these two experiments, then one expects them to be using a representation that functions as an intervening variable, and this predicts a failure of independence. The theory-of-mind hypothesis asserts that chimps develop beliefs that distinguish the epistemic states of knower and guesser in the first experiment, and then use those beliefs to distinguish the epistemic states of knower and guesser in the second. This is the (IV–1, 1, 2) model, which fails to gain support when it is tested against the (N–1, 2) hypothesis.

11 Concluding comments

By now, virtually all psychologists recognize that it makes no sense to reject categorically and unconditionally the postulation of internal states. The question remains, however, of when intervening variables should be postulated. In this essay, I have argued that this problem about psychological theorizing can be investigated in the more general setting of black box inference. The observed association of inputs and outputs should take on different patterns, depending on whether intervening variables exist and how they are structured. This idea—that model selection can be data-driven—contrasts sharply with the idea that mentalistic hypotheses are to be favoured because they are more parsimonious than their competitors. It isn’t that parsimony is never relevant to this problem. Rather, my goal has been to show how model selection can be based on observations; in this circumstance, parsimony as a consideration can be razored away (Sober [1994a, 1998]).
In comparing a no-intervening-variable model with a model that postulates an intervening variable, I interpreted the former as advancing a claim of probabilistic independence and the latter as claiming that a certain relation of probabilistic dependence should obtain. Within this framework, the hypothesis of independence has the status of a null hypothesis, regardless of which total model has fewer adjustable parameters. The operative maxim seems simply to be that we should postulate an intervening variable, or not, depending on the character of the data. If this is the right principle, then the issue of which model is more parsimonious in the sense of containing fewer adjustable parameters really is besides the point.

I do not claim that the methodology I have described is a panacea. I do claim that it is useful, both because it clarifies what scientists are already doing and because it suggests a test that might be deployed in future investigations. Many psychologists and philosophers have had the intuition that response generalization and stimulus generalization are extremely important phenomena when it comes to justifying the postulation of intervening variables. The present paper explains why this is so.

It is natural to interpret the results of Wickens [1938] and of Povinelli et al. [1990] in terms of an intervening-variable model. In both cases, experience in an initial experiment seems to raise the probabilities of two behaviours and does so by changing some internal state that is a cause of each. We have seen that the second conjunct in this interpretation is not supported by the data gathered. This means that if you believe that there is an intervening variable in either study, this belief is not based on the data gathered but merely reflects some prior conviction that you bring to the experiment. One virtue of the

Some intervening-variable models are more parsimonious than their no-intervening-variable competitors, but sometimes, the reverse is true. For example, (IV—3, 1, 3) is more parsimonious than (N—3, 3) when all events are dichotomous, but (IV—1, 1, 2) is less parsimonious than (N—1, 2). However, it is a separate question whether claims of independence are more parsimonious than claims of nonindependence.

Hypotheses of independence are always more parsimonious. For example, if Pr(E1|C) = a and Pr(E2|C) = b, then the four probabilities Pr(±E1 & ±E2|C) have the following form:

<table>
<thead>
<tr>
<th>E2</th>
<th>not-E2</th>
</tr>
</thead>
<tbody>
<tr>
<td>E1</td>
<td>ab + c</td>
</tr>
<tr>
<td>not-E1</td>
<td>(1 - a)b - c</td>
</tr>
</tbody>
</table>

The hypothesis that C renders E1 and E2 conditionally independent means that c = 0; if independence fails, then c is an adjustable parameter constrained to be greater than 0. The hypothesis of independence thus has one fewer adjustable parameter. Notice that a, b, and c can be estimated from observed frequencies, which means that the two hypotheses are both identifiable. The framework of Akaike [1973] therefore applies (Forster and Sober [1994]) and explains why parsimony matters in this inference problem; parsimony is relevant to estimating how predictively accurate the two hypotheses will be.

In standard Neyman–Pearson statistics, the hypothesis of probabilistic independence is treated as the null hypothesis. Although this framework makes no explicit reference to parsimony, it in fact gives weight to parsimony by saying that the null hypothesis should be rejected if and only if Pr(Data|Null) is lower than some specified value.
probabilistic methodology of black box inference is that it helps separate interpretations that are licensed by the data from interpretations that, if justified, are justified in some other way.

The methodology I have described provides a suggestion that can be used in the design of experiments. Suppose that all the chimps in Povinelli et al.'s study had success rates of 95% at the end of the first experiment and that all had 90% success rates at the start of the second. This might seem to be strong evidence for the theory-of-mind hypothesis, but, as we have seen, one will not be able to reject the null hypothesis of independence if all subjects behave the same. To confirm the intervening-variable model, subjects must behave differently; in particular, subjects with higher than average scores at the end of the first experiment should also have higher than average scores at some point in the second.

I suspect that work in this and other areas of cognitive science may be influenced, if only implicitly, by the following line of reasoning. The goal is to determine whether the individuals in a given species have such-and-such a trait, on the assumption that the trait does not vary within species; for example, when the question is posed of whether chimps have a theory of mind, the assumption is that if one chimp has a theory of mind, they all do. This is taken to mean that a good experiment that tests for the presence of the trait should have all subjects behaving the same, or nearly the same. (Compare: if dust allergy is either universally present or universally absent in a group of subjects, then a good test for the allergy should come out the same, or nearly the same, for all subjects.) As a result, variation in performance is a sign that one has devised a bad experiment.

This line of reasoning offers bad advice, because it leads one to construct experiments that cannot provide evidence for rejecting the null hypothesis of conditional independence if it is false. But there is a second defect in the imagined line of reasoning. If the chimps do not behave the same in an experiment, this does not mean that they differ with respect to their capacities; perhaps the ability to formulate second-order intentional representations is universal, but only some of the chimps in the experiment actually do so. The allergy analogy is misleading in this respect. An experiment in which subjects vary in their behaviour is needed if one is to obtain evidence for the existence of an intervening variable; however, this variation is consistent with the assumption that one is studying a species-wide capacity.

There is more to mentalistic hypotheses than the postulation of intervening variables. Attributions of beliefs and desires not only claim that intervening variables exist, but assign them a semantic content. The methodology I have described does not provide ground rules for this second task. Those working in cognitive ethology and related areas often appeal to Morgan's Canon and to other principles that counsel conservatism or parsimony (Dennett [1983];
Sober [1998]). The advice generated by these principles may seem intuitive, but the logical foundations of the principles themselves are less than obvious. There presently exists a very large gap in our understanding of how attributions of mental states should be evaluated. None the less, the argument of the present paper suggests a kind of baseline: if an intervening-variable model cannot be justified, it is questionable whether one should spend time arguing about exactly what semantic content those supposed intervening variables possess.

As important as this reality check on mentalism is, the methodology of black box inference goes further. We saw earlier that the same principles that distinguish one intervening variable from none also provide advice concerning how many layers of intervening variables should be introduced. This may be relevant to psychological investigations in which different hypotheses postulate different amounts of internal processing. For example, discussions of theory of mind typically pit a hypothesis that attributes second-order intentionality against a more 'conservative' hypothesis that attributes first-order intentionality only. Most philosophers agree that an individual who has second-order intentionality also must have first-order intentionality, but that the converse entailment does not hold. If a chimp can form the belief that a trainer knows where the food is, then the chimp also must be able to form beliefs about the location of food; the reverse claim is not correct, however. This is often regarded as a logical point about the concept of belief—if it is possible for S to believe that O believes that p, then it is possible for S to believe that p. Whether or not this principle is 'logical' (or a priori), it is synchronic; organisms that have second-order intentionality at a given time also have first-order intentionality at that time.

I want to propose a supplement to this synchronic claim. My proposal is diachronic and causal: second-order intentional states are caused by first-order intentional states. If a chimp believes that a trainer knows where the food is, this second-order belief is caused by the chimp’s having various first-order beliefs about the trainer (e.g. that he was near the food when it was hidden). However, first-order beliefs need not be caused by second-order beliefs (though they may be in some instances). If this is right, then a hypothesis of second-order intentionality may postulate more layers of intervening variables than a hypothesis that asserts first-order intentionality only. If every cause described by these hypotheses is said to have more than one effect, then the hypotheses will make different predictions. The perspective of black box inference therefore suggests a way in which observations may eventually be able to resolve the question of whether chimps have a theory of mind.

It would be interesting to compare the conditional-independence test that I have proposed with the more familiar method of testing processing models by considering reaction times; see Sternberg [forthcoming] for discussion. However, this question is beyond the scope of the present paper.
More than a few philosophers have maintained that the special subject matter of psychology engenders unique methodological problems for that science. However, before one concludes that psychology is an idiosyncratic discipline, it is worth trying to identify inference problems in which psychology differs from other sciences in content, not in form. Black box inference is a case in point.

Acknowledgements

I first want to thank Branden Fitelson for his detailed comments on multiple drafts of this paper. I also am grateful to Colin Allen, Ron Amundson, Martin Barrett, Marc Bekoff, Len Berkowitz, Berent Enç, Malcolm Forster, Clark Glymour, Dan Hausman, Celia Heyes, David Lorvick, Greg Mougin, Robert Pennock, Daniel Povinelli, Larry Shapiro, Sam Sober, Chris Stephens, Kim Sterelny, and Ann Wolfe for useful discussion.

Department of Philosophy
University of Wisconsin, Madison
Madison, WI 53706
USA

References


