Explaining the Brain

Mechanisms and the Mosaic Unity of Neuroscience

Carl F. Craver

CLARENDON PRESS · OXFORD

OXFORD
UNIVERSITY PRESS
Great Clarendon Street, Oxford OX2 6DP
Oxford University Press is a department of the University of Oxford.
It furthers the University's objective of excellence in research, scholarship,
and education by publishing worldwide in
Oxford New York
Auckland Cape Town Dar es Salaam Hong Kong Karachi
Kuala Lumpur Madrid Melbourne Mexico City Nairobi
New Delhi Shanghai Taipei Toronto
With offices in
Argentina Austria Brazil Chile Czech Republic France Greece
Guatemala Hungary Italy Japan Poland Portugal Singapore
South Korea Switzerland Thailand Turkey Ukraine Vietnam
Oxford is a registered trade mark of Oxford University Press
in the UK and in certain other countries
Published in the United States
by Oxford University Press Inc., New York

British Library Cataloguing in Publication Data
Data available

Library of Congress Cataloging in Publication Data
Craver, Carl F.
Explaining the brain : mechanisms and the mosaic unity of neuroscience / Carl F. Craver.
p. ; cm.
Includes bibliographical references and index.
ISBN–13: 978–0–19–929931–7 (alk. paper)
ISBN–10: 0–19–92993 1–5 (alk. paper)
1. Neurosciences—Philosophy. 2. Brain—Philosophy. I. Title.
[DNLM: 1. Neurosciences. 2. Philosophy. WL 100 C898e 2007]
QP356.C73 2007
612.8—dc22 2006103230

Typeset by Laserwords Private Limited, Chennai, India
Printed in Great Britain
on acid-free paper by
Biddles Ltd., King's Lynn, Norfolk

ISBN 978–0–19–9299317

10 9 8 7 6 5 4 3 2 1

Detailed Contents

motivation. For in Glennan's account, the most pressing questions for a normatively adequate account of causation are stipulated as features of "direct causal laws" at the fundamental level. As a consequence, Glennan neither ameliorates Hume's worries nor satisfies (E1)–(E5). In Section 5, I show how the manipulationist approach satisfies (E1)–(E5) without stipulation and without descending into fundamentalism.

5. Manipulation and Causation

I dwell at some length on the mechanical and transmission approaches to causation because each is associated with mechanistic explanation and because each can be used (intentionally or not) to support explanatory fundamentalism. The mechanical view grounds higher-level causal relations in lower-level mechanisms, a grounding process that ends, if ever, only in the most fundamental causal laws. The transmission account is more explicit in its association between causation and properties found only at the most fundamental levels (that is, conserved quantities). The fact that neither of these views provides an adequate account of causation—and in particular, that each struggles to provide an account of causal relevance and negative causation—weakens the attraction of fundamentalism.

To repeat a central theme: causal relevance, explanation, and control are intimately connected with one another. This is particularly true in biomedical sciences, such as neuroscience, that are driven not merely by intellectual curiosity about the structure of the world, but more fundamentally by the desire (and the funding) to cure diseases, to better the human condition, and to make marketable products. The search for causes and explanations is important in part because it provides an understanding of where, and sometimes how, to intervene and change the world for good or for ill. This connection between causation, explanation, and control is also reflected in the procedures that neuroscientists use to test explanations. These tests involve not only revealing correlations among the states of different parts of a mechanism but, further, intervening in the mechanism and showing that one has the ability to change its behavior predictably. More explicitly: to say that one stage of a mechanism is productive of another (as I suggest in Machamer et al. 2000; Craver and Darden 2001), and to say that one item (activity, entity, or property) is relevant to another, is to say, at least in part,

that one has the ability to manipulate one item by intervening to change another. More concretely, to say that LTP is caused by tetanic stimulation is to say that one can potentiate a synapse by tetanizing it.

In embracing this view, I rely closely on James Woodward's account of the role of invariance in explanation (see, especially, Woodward 1997, 2000, 2002, 2003; and Woodward and Hitchcock 2003a, 2003b). Woodward is not especially concerned with neuroscience; however, he is concerned with developing an account of causation adequate for explanations that involve mechanistically fragile and historically contingent generalizations. Woodward (2002) shows how his account of causal relations might be fitted into an account of mechanisms, and Glennan (2002) has followed him in this idea. In this section, I build on that idea by showing how the manipulationist account of causal relevance can satisfy (E1)–(E5). I also show how it can accommodate negative causation.

Woodward's view is currently the most defensible and readable exposition of the manipulationist tradition in thinking about causation both in philosophy (see, for example, Collingwood 1940; von Wright 1971) and in statistics (Cook and Campbell 1979; Freedman 1997). Related ideas appear in Pearl's (2000) notion of a "do operator," the notion of an intervention by Spirtes et al. (1993), and Glymour's (2001) idea of surgically intervening into a causal graph. The central idea is that causal relationships are distinctive in that they are potentially exploitable for the purposes of manipulation and control. More specifically, variable X is causally relevant to variable Y in conditions W if some ideal intervention on X in conditions W changes the value of Y (or the probability distribution over possible values of Y). In the context of a given request for explanation, the relationship between X and Y is explanatory if it is invariant under the conditions (W) that are relevant in that explanatory context. Now I consider the different components of this basic statement.[21]

Woodward construes X and Y as variables, that is, as determinables capable of taking on determinate values. Although this is a common way of speaking in some areas of science and statistics, philosophers have generally

[21] Again, I do not offer this account as a reductive analysis of causation. It would clearly be circular, given that intervention is an ineliminably causal concept. Instead, my account is intended as a necessary condition to be met by relationships of causation and to be explained by any satisfactory metaphysics of causation. Lewis's view of causation, for example, ably captures many of the crucial features of this necessary condition (see Woodward and Hitchcock 2003a for a discussion).

preferred other relata in their accounts of causation. Davidson (1969) and many other philosophers, for example, describe causation as a relationship between events. Salmon (1984, 1998) and Dowe (2000) describe it as a relationship among processes. Others describe it as a relationship among objects, facts, and contrasts. Each of these ways of speaking and thinking about causation can be translated without loss into talk of variables. For example, talk of event and object causation can be translated into talk of a variable that can take on two values {E occurs/is present, E does not occur/is not present}. Talk of causation among processes can be translated by assigning variables to the features of a process or to the magnitudes of the conserved quantities.[22] Similar translations can be made for the other ways of thinking about causation. To view causal relevance as a relationship among variables allows one to consider cases in which the variable may take on any value in a continuum (for example, a dose), to make relative assessments of causal efficacy along that continuum (for example, a dose-response relation), and to consider cases in which there are sharp discontinuities in the effect between one portion of the continuum and another (threshold events, such as action potentials).[23]

The term "intervention" denotes, roughly, a manipulation that changes the value of a variable. It is helpful to think of interventions as well-designed experimental interventions. However, one must not think of manipulations as exclusively the products of human agency. When a stroke damages a brain region, this counts as an intervention on that brain region's functioning. When a meteor strikes the moon, it intervenes in the moon's environment.

The manipulationist view of causal relevance requires that the relationship between X and Y must be *potentially* exploitable for the purposes of manipulating Y in conditions W. One need not actually be able to manipulate X. One might not know how to intervene on X, one might not have the tools, or X might be too small, too big, or too far away for human intervention. Many believe, for example, that a spatial map in the

[22] Those who think of causation as involving activities can make use of the fact that activities have precipitating conditions or enabling properties (that are necessary for or conducive to the occurrence of the activity) and termination conditions or signatures (that is, effects). One can then apply the strategy just described for causation among events, objects, or properties. See Darden and Craver (2002).

[23] As I note above and discuss further in Chapter 6, a contrastive formulation is even more perspicuous. It is a variable X's having one value (rather than some other value) that causes the effect to occur (rather than some alternative). (See Dretske 1977; Hitchcock 1996.)

hippocampus is causally relevant to the ability of rodents to navigate their environments (as argued by O'Keefe and Dostrovsky 1971; O'Keefe and Nadel 1978; Wilson and McNaughton 1993). They believe this in spite of the fact that neuroscientists currently lack the ability to drive a rat through a novel maze by manipulating its spatial map. The ability to do so would no doubt be convincing evidence that the hippocampus is involved in navigation, but this evidence is not required to know that there is a causal relation. What matters is that there is a relationship between X and Y that can possibly be exploited to change Y by changing X, even if no human can or will ever be able to so exploit it. It is a very interesting question how (and how much) we can manage to learn about the causal structure of the world in cases where we cannot intervene in this way. This question is best answered through a detailed look at specific experimental practices in neuroscience. I do not pursue such a detailed investigation in this book (see, for example, Bogen 2001; Bechtel forthcoming). I focus instead on more abstract and general features of the evidence required to establish causal claims.

An *ideal* intervention I on X with respect to Y is a change in the value of X that changes Y, if at all, *only via* the change in X. More specifically, this requirement implies that:

(I1) I does not change Y directly;
(I2) I does not change the value of some causal intermediate S between X and Y except by changing the value of X;
(I3) I is not correlated with some other variable M that is a cause of Y; and
(I4) I acts as a "switch" that controls the value of X irrespective of X's other causes, U. (Adapted from Woodward and Hitchcock 2003a)

These restrictions on ideal interventions are represented graphically in Figure 3.4. Unidirectional arrows represent causal relations, bidirectional dotted arrows represent correlations, and bars across arrows represent a restriction against the represented relation. In this figure, an intervention changes the value of X, surgically removing other causal influences, U, on X (I4). This intervention produces a change in Y that is not mediated directly (I1), by affecting an intermediate variable, S (I2), or by being correlated with some other variable, C, that can change the value of Y

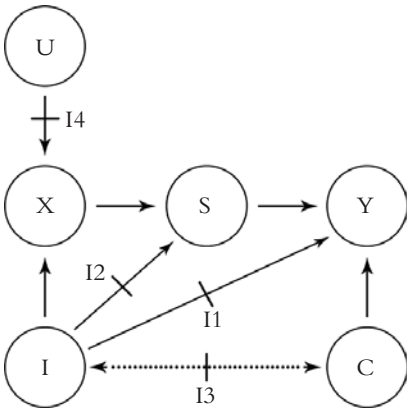

Figure 3.4. An ideal intervention on X with respect to Y*

* Solid arrows represent causal relations; dotted arrow represents correlations; hashes represent the absence of the cause or correlation

(I3). Note that conditions (I1)–(I4) represent the kinds of control that are routinely used and required to test causal and explanatory claims.

The focus on ideal interventions will give rise to objections that experimental situations are often in many ways non-ideal. This is true, and it is an important insight about our epistemic situation with respect to the causal structure of the world. More work remains to be done to say how one can learn about the causal structure of the world if criteria (I1)–(I4) are relaxed, removed, or replaced in order to more accurately describe the complex epistemic situation in which most experimentalists work. The best inroad into that discussion, it seems to me, is to work first on the clear cases and then to see how (and if) the account can be adjusted so that it can regiment non-ideal experimental situations.

Consider the LTP example again. When is it appropriate to assert that a tetanus in the pre-synaptic cell is causally relevant to LTP? One might establish this relationship experimentally by intervening into the pre-synaptic cell, delivering a tetanus, and observing subsequent changes in the strength of the synapse. An experimenter could intervene by injecting current into the cell, by creating an electrical field in a population of pre-synaptic cells, by applying neurotransmitters to the pre-synaptic cell, or by allowing a population of cells to enter its normal burst cycle. What matters is that the intervention makes the pre-synaptic cell fire rapidly and

repeatedly. Suppose that one performs such an intervention and observes a subsequent increase in the strength of the synapse. Such a finding would not warrant belief in the claim that the tetanus is causally relevant to synaptic strength. It is possible that the intervention strengthened the synapse for reasons having nothing to do with the tetanus. Perhaps merely breaking the cell membrane or inserting an electrode into a population of neurons can strengthen synapses (in violation of (I1)). Or perhaps inserting the electrode changes features of neurotransmitter release (in violation of (I2)). Or perhaps one inserts electrodes only when one has the cells in a particular bath solution, and the bath solution strengthens synapses (in violation of (I3)). Or perhaps the injected current is swamped by input from other neurons into the cell (in violation of (I4)). If any of the conditions (I1)–(I4) fails in an experimental protocol, the observed changes in synaptic strength would not be good evidence of a causal relationship between the tetanus and the changes in synaptic strength. When one asserts a causal relevance relation between the firing rate of the pre-synaptic cell and the strength of a synapse, one asserts when one alters the firing rate of the cell in specified ways using an ideal intervention, then one either strengthens the synapse or changes the probability that the synapse would be strengthened.

Each of the activities in the LTP mechanism can be described in the same way. Neuroscientists believe that glutamate opens NMDA receptors because they open when glutamate is applied, but not (or not to the same extent) when isotonic saline or some other neurotransmitter is applied, and not when the binding site for glutamate has been blocked or altered. They are convinced that Mg^{2+} blocks the flow of Ca^{2+} into the post-synaptic cell because they can manipulate Ca^{2+} levels in the cell by changing the concentration of Mg^{2+} or by manipulating the electrical potential that holds Mg^{2+} ions in the NMDA receptor's pore. They are convinced that depolarizing the post-synaptic cell is relevant to the eventual occurrence of LTP because they can keep everything else the same and eliminate LTP simply by clamping the voltage of the post-synaptic neuron at rest. Experiments of this sort show neuroscientists what can manipulate what. On the further assumption that such manipulations are relevantly similar to changes occurring in the brain under the conditions in question, neuroscientists can assume that natural interventions (that is, those not wrought by human hands) produce similar changes in the brain.

5.1 Invariance, fragility, and contingency

The explanatory generalizations describing these causal relevance relations are stable, or as Woodward says invariant, though not necessarily—or even usually—universal. To say that a generalization is stable is to say that the specified relation between the cause variable and the effect variable holds under a (generally nonuniversal) range of conditions. The conditions under which a generalization might be stable include *stimulus conditions, intermediate conditions*, and *background conditions*. Stimulus conditions include conditions explicitly represented as independent variables in the description of the relationship; in the case of X →Y, the stimulus condition is X. The relationship need not be stable across all stimulus conditions. Outside of a normal range of stimulus conditions, the stimulation might have no effect, might weaken the synapse, or might simply damage the cells. The generalization might also be more or less stable under a range of values for the variables intermediate between X and Y, such as Ca^{2+} concentration and Mg^{2+} concentration. Finally, the relationship holds only under a range of background conditions, such as temperature, pH, and available energy. Stable causal relations in neuroscience, in other words, do not hold under all conditions but only under a narrow range of conditions.

The idea that a relationship between variables must be stable to be explanatory is also weaker than the requirement of "contextual unanimity" found in many accounts of causation (for example, Cartwright 1983; Eells 1991; Skyrms 1980). The requirement of contextual unanimity demands, roughly, that if X causes Y, then the relationship between X and Y holds in all contexts. This requirement is too strong for the causal relations in neuroscience precisely because these causal relationships often depend crucially upon the absence of counteracting causes, on the absence of interaction effects, and on background conditions within relatively circumscribed ranges (see Glennan 1997). In contrast to the contextual unanimity requirement, the manipulationist approach allows explanatory generalizations to vary considerably in their stability or invariance and requires only that the generalization should be stable in the conditions relevant to a particular request for explanation (see below).

The fact that generalizations can be more or less stable and still be explanatory is useful for dealing with the fact that causal generalizations in neuroscience are limited in scope, mechanistically fragile, and historically

contingent. Causal relations need not be universal to be explanatory, nor need they be unrestricted in scope, nor need they lack any reference to particulars. All that matters is that there is some stable set of circumstances under which the variables specified in the relation exhibit the kind of manipulable relationship sketched above. Mechanistically fragile generalizations are invariant over a range of values for the stimulus variable, the intermediate variables, and the background conditions. Furthermore, the fact that the relationship is historically contingent, and so in some sense unnecessary, makes no difference to whether the relation is explanatory here and now. Na^+ channels produce action potentials today even if no creatures produce action potentials that way 20 million years from now. What matters, again, is that there exists a range of conditions under which one can reliably manipulate the effect variable by intervening to change the cause variable.

Which are the relevant conditions for assessing the stability of a generalization? There is no general answer to that question. Woodward often confines his attention to changes in the values of the variables appearing in the statement of the causal relevance relation. However, this requires at once too much and too little. It requires too much because, as just noted, such relationships might break down under extreme values of the variables appearing in the statement of the relation. It requires too little because, although neuroscientists are often interested in physiologically relevant conditions (that is, the conditions found in intact and healthy organisms), they are just as often interested in disease states in which the stimulus, intermediate, and background conditions are abnormal or pathological. Sometimes they are interested in background conditions well outside the physiological range, as when they try to explain highly contrived experimental effects, to design drugs to interact with the CNS, or to commandeer some part of the CNS for their own purposes. The appropriate range of conditions in which a causal generalization must be stable thus depends crucially upon one's explanatory interests. This does not mean that the causal relations are interest-relative. The causal relevance relations under different ranges of conditions are objective features of the world. However, which of those objective relations is relevant depends on what you are trying to explain.

5.2 Manipulation and criteria for explanation

According to the manipulationist account, explanatory texts describe relationships between variables that can be exploited to produce, prevent,

or alter the *explanandum phenomenon*. Merely being able to manipulate a phenomenon, of course, is not sufficient to explain it. People made babies long before they understood how DNA works. But the wider the range of possible manipulations, and the deeper one's knowledge of how such manipulations change the *explanandum phenomenon*, the more complete is the explanation. As Woodward puts it, a good explanation allows one to answer a range of "what-if-things-had-been-different questions" (w-questions, for short). Deep explanatory texts (or models) provide the resources to answer more questions about how the system will behave under a variety of changes than do shallow explanatory texts. The answers to such questions are evaluated experimentally according to the standards described above.

The manipulationist view readily satisfies criteria (E1)–(E5). Consider mere time-courses (E1). The ability to lay down long-term memories invariably appears after the development of the primary sexual characteristics, but (so far as I know) the latter is explanatorily irrelevant to the former. In contrast, delivery of a rapid and repeated stimulus to the pre-synaptic cell is explanatorily relevant to the entry of Ca^{2+} into the post-synaptic cell. The difference, according to the manipulationist account, is that one could not manipulate the ability to lay down long-term memories by intervening to change the development of primary sexual characteristics (so far as I know), but one can manipulate the tetanus to change the concentration of Ca^{2+} in the post-synaptic cell. This way of dealing with the difference between causation and regular succession has clear advantages over both regularity-based accounts of causation and certain counterfactual views. Both of these alternative views of causation treat at least some cases of regular temporal succession as cases of causation. This is because the values of the two variables, X and Y, are constantly conjoined (*ex hypothesi*) such that whenever the first variable occurs, the second does as well. One could then infer that if X takes a particular value, then Y will take the corresponding value. Nonetheless, it is not the case that one could change Y by intervening to change X. In cases of this sort, the relationship between X and Y supports what Lewis (1979) calls "backtracking counterfactuals," but, as Lewis notes, such counterfactuals are not explanatory. The manipulationist-based approach instead requires causal regularities to fulfill a more demanding requirement, namely that if X is set to x in accordance with (I1)–(I4), then Y will take on the value f(x). This kind of statement

is tested in controlled experiments. Relations that meet this requirement allow one to answer w-questions.

The same strategy can be used to show why causal explanations tend to run from earlier to later (E2). The reason is that, at least in all known cases in neuroscience, one cannot change the past by intervening in current states of affairs. No matter what one does to the pre- or post-synaptic neuron now, one will not change the way that it behaved yesterday. There is no need to *insist* that all causes precede their effects on metaphysical grounds. There are still debates about whether backwards causation is possible in physics (see, for example, Price 1996). Were such a relationship to be demonstrated (using the sorts of ideal experimental manipulations discussed above), one would be justified in asserting that past events can be caused by future events and in asserting that at least in some cases one needs to appeal to future events to explain the past. However, there have been no such demonstrations in neuroscience, and this helps to explain the presumption that explanations in neuroscience are temporally asymmetrical.

Constraint (E3) is that two effects of a common cause do not explain one another in spite of the fact that the occurrence of one allows us to infer the occurrence of the other. Suppose that one pre-synaptic neuron (A) synapses upon two unconnected post-synaptic neurons, N_1 and N_2; that stimulating A reliably causes N_1 and N_2 to fire action potentials; and (for simplicity) that N_1 and N_2 are quiescent in the absence of activity in A. Let X be a variable representing the electrical activity of N_1 with the values {firing, not firing} , and let Y be a like-valued variable representing the electrical activity of N_2. Under these suppositions, one could reliably infer the value of X from the value of Y and vice versa because N_1 and N_2 always fire in tandem. That is, there is a robust regularity between X and Y that sustains certain backtracking counterfactuals. Were X to take the value {firing}, then Y would take the value {firing}. And if Y were to take the value {not firing}, then X would take the value {not firing} as well. However, one could not change X's activity by intervening directly to change Y. Nor could one change Y's value by intervening directly to change X. The regularities here do not satisfy requirement W. Examples such as this generalize: if the relationship between two variables is merely a correlation, then one will not be able to manipulate one variable by intervening to change the other. If the two are causally related, then one can manipulate one of them by manipulating the other.

The manipulationist approach also sorts relevant from irrelevant properties and interactions, as required by criterion (E4). (I extend this basic model considerably in Chapter 6 to address issues of nonfundamental causal relevance). To begin with the parson and his micropipette, while it may be true that all pyramidal cells blessed with holy water produce LTP when tetanized, the holy water is irrelevant. One can establish this by intervening in the above sense to remove the blessing, or to change the blessing to a curse, while leaving everything else the same. If one finds that such interventions have no effect on the occurrence (or incidence) of LTP, then one should conclude that the blessing is irrelevant to LTP. Of course, experiments are rarely so clean in the real world. In the history of LTP research, for example, it has been very difficult to determine which of the myriad interactions among intracellular molecules are relevant to the occurrence of LTP (see, for example, Sanes and Lichtman 1999). Part of the reason that these relevance relationships have been so difficult to disentangle is that the intracellular molecular cascades are so complex and causally interwoven that it is difficult to perform the sorts of ideal interventions described above. It is complex, in practice, to determine that one's intervention acts only on the target variable X, and that the intervention changes Y only via X and not through a host of myriad other connections. But these practical difficulties, which are part of what make science challenging and rewarding, do not impugn the overall idea that what one ideally wants to establish is precisely such well-controlled relationships of manipulability.

The final criterion, that the account of causation should allow for improbable effects (E5), requires only a slight modification of the basic argument scheme applied to (E1)–(E4). Many of the causal relationships posited in neuroscience are probabilistic. Tetanizing a pre-synaptic cell produces LTP only 50 percent of the time (with current techniques). If X and Y are only probabilistically related, then any particular intervention to change X might have no effects on Y. As remarked above, what the manipulationist account requires in such cases is that manipulating X changes the probability distribution over possible values for Y. For example, depolarizing the neuron should change the probability that the Na^+ channel will open or that the synapse will be potentiated. In neither case is it required that manipulating X makes Y probable (that is, $p(Y \mid X) > 0.5$). The probability of Y might be quite low even under the maximally

effective manipulations of X. Indeed, this matches precisely the way that researchers assess stochastic relationships in neuroscience and elsewhere.

One last point requires emphasis. Nothing in this view of causal relevance makes reference to a privileged level at which all causes act or at which all relevant causes are located. Variables can be fundamental (spin, charm) or nonfundamental (socio-economic status, priming, inflation). All that matters is that they exhibit the patterns of manipulability discussed above.

5.3 Manipulation, omission, and prevention

A final promising feature of the manipulationist approach to causal relevance is that it accommodates causation by omission and prevention (see Woodward 2002). In cases of omission, such as when the absence of an attractive force allows the Mg^{2+} ion to float out of the NMDA receptor channel, what matters is not the transmission of marks or conserved quantities from the beginning of this mechanism to the end, but rather the fact that one can prevent the Mg^{2+} ion from floating out of the channel by polarizing the cell. Likewise in cases of prevention, such as when the Mg^{2+} ion blocks the channel and thereby prevents Ca^{2+} from entering the cell, what matters is not an exchange of conserved quantities between the Mg^{2+} ion and the non-increase in Ca^{2+} (for there can be no such exchange), but rather the fact that by manipulating the putative cause (positive or negative), one can make a difference in the putative effect (positive or negative).

The ability of the manipulationist account of causal relevance to satisfy (E1)–(E5) and to accommodate cases of negative causation is directly tied to the ability of such generalizations to answer w-questions. This ability provides the kind of rich information about the *explanandum phenomenon* that is typically required of a good explanation. When one knows the relations of manipulability, one can say which interventions make a difference to the *explanandum* and which do not (for example, mere temporal predecessors, temporal successors, irrelevant properties, and the like). In cases where interventions do make a difference, knowing these relations allows one to predict how the *explanandum phenomenon* will be different under a variety of conditions. There is a strong appeal in this connection given that one way to test one's understanding of a phenomenon (as any good test-writer knows) is to test whether someone can say how it will change in novel conditions.