Moving from Levels & Reduction to Dimensions & Constraints

David Danks (ddanks@cmu.edu)
Department of Philosophy, 135 Baker Hall, Carnegie Mellon University
Pittsburgh, PA 15213 USA

Abstract

Arguments, claims, and discussions about the "level of description" of a theory are ubiquitous in cognitive science. Such talk is typically expressed more precisely in terms of the granularity of the theory, or in terms of Marr's (1982) three levels (computational, algorithmic, and implementation). I argue that these ways of understanding levels of description are insufficient to capture the range of different types of theoretical commitments that one can have in cognitive science. When we understand these commitments as points in a multi-dimensional space, we find that we must also reconsider our understanding of intertheoretic relations. In particular, we should understand cognitive theories as *constraining* one another, rather than reducing to one another.

Keywords: Level of description; Marr; Philosophy of cognitive science; Reduction; Intertheoretic constraint

Limitations of Levels

It is customary within science to talk about our theories as falling at different "levels of description": biology is at a higher level of description than chemistry, which is itself at a higher level than physics. Moreover, talk of levels is not restricted to the relationships between these large-scale domains of science; a sub-symbolic model of causal cognition can be said to be at a lower level of description than some symbolic model of the same cognition or behavior.

"Levels talk" is particularly widespread in the cognitive sciences (as noted by many authors, such as Bechtel, 1994; Bickle, 1998; Marr, 1982). The proliferation of talk about levels is quite unsurprising, given the many different methodologies used to develop theories of human behavior and cognition. At the same time, exactly what is meant by a "level" is often left somewhat vague. Levels of description are sometimes identified with the ontological granularity of a theory, where its level is determined (largely) by its objects. This characterization misses important distinctions, however, such as the difference between a rational analysis that says how one should act, and a process model that describes the cognitive mechanisms generating behavior.

One of the most precise characterization of levels in cognitive science—and certainly the most influential such characterization—was given by Marr (1982), and captured this key distinction. Marr's three levels characterize information-processing devices in general, and processes in the human mind more specifically. The computational level identifies the input and output of the process, as well as constraints on the types of computation done on the input to get the output. The algorithmic level (also called the representation level) specifies an implementation of the computational theory, as well as the representation of the input and output of the process. Finally, the implementational level describes the physical realization of the representation and the algorithm.

Roughly speaking, the computational level specifies what problem is being (appropriately) solved; the algorithmic level explains how it is solved; and the implementational level gives the details of the physical substrate that does the solving. As a concrete (non-cognitive) example, we can understand a word-processing program as (i) a process for entering, editing, and rendering text documents (the computational level); (ii) a bunch of lines of code that produce the appropriate behavior (the algorithmic level); or (iii) changes of 1's and 0's in the internal memory registers of the computer (the implementational level).

As a more cognitive example, consider the problem of learning causal structure from observational data (e.g., Cheng, 1997; Griffiths & Tenenbaum, 2005). A computational-level model of this problem would characterize the relevant inputs (case-by-case observations or a summary of a sequence of such cases), the output that should result given such input (a representation that can be used for causal inference, decision-making, explanation, etc.), and any relevant cognitive constraints (though in practice, computational-level models rarely incorporate such constraints). An algorithmic-level model would characterize the internal representations and cognitive processes by which we humans happen to solve this challenge. And an implementation-level model would show how the relevant computations are performed in particular brain regions (e.g., frontal cortex as suggested by Fletcher, *et al*., 2001 or Satpute, *et al*., 2005).

Marr's three levels were a significant advance in part because they are based on the recognition that the mathematical or computational specification of a cognitive theory significantly underdetermines the commitments that are implied by it. A Bayesian model of causal learning could, for example, be at the computational or algorithmic level, depending on the intended interpretation of the terms in the model. Moreover, these differences in interpretation (and so commitments) can matter: whether some experiment or behavioral measure is a test of a model depends in part on the commitments of that model.

Marr's levels were also intended to help show that there can be distinct models of the same phenomenon that are *not* competitors. That is, models M_1 and M_2 can be incompatible (whether mathematically or ontologically) and yet both be correct as long as they are at different levels. For example, Bayesian and associationist models of causal learning are mathematically incompatible—they posit different representations and different learning processes—but can

both be correct if one is at the computational level and the other is at the algorithmic level (Danks, Griffiths, & Tenenbaum, 2003; Griffiths & Tenenbaum, 2005).

Unfortunately, Marr's levels suffer from at least two significant flaws. First, and more importantly, they assume that multiple distinct aspects of theoretical commitment must vary together, rather than being able to vary independently (see also McClamrock, 1991). For example, suppose model M_1 is a standard computational-level model of human causal learning: it characterizes the relevant inputs and shows which (behavioral) outputs would solve the causal learning task, all while being agnostic about the underlying representations and processes.

Now consider M_2 that is mathematically identical to M_1, but which claims only that people *do* generate this (behavioral) output, not that this behavior is how people *should* solve the causal learning task. That is, M_2 is a relatively standard instrumentalist model that characterizes the human behavior without explaining precisely how or why it is generated. M_2 is not a computational-level model, as it does not explain why people act as they do (i.e., one of the putative hallmarks of a computational-level model). At the same time, M_2 is not an algorithmic-level model, as it does not characterize the underlying representations or cognitive processes. There thus does not appear to be any place to put M_2 in the standard three Marr levels.

More generally, Marr's three levels force three different dimensions of variation in theoretical commitment—extent of realism, tightness of approximation, and (importance of) closeness to optimality (all discussed in the next section)—to change in lockstep when they can, in practice, vary relatively independently. This observation points towards the second concern about Marr's levels: namely, each of these dimensions has many more than just three levels, as theories can differ (in their commitments) in relatively fine-grained ways. Marr's levels are sometimes helpful for providing a quick characterization of the commitments of some theory, if the theory happens to fit one of those templates. But in general, we need a subtler characterization of the types of theoretical commitments we can have for a given cognitive model.

Dimensions of Variation in Commitments

In this section, I consider in more detail these three dimensions of variation in one's theoretical commitments. At the end, I show how we can use these dimensions to better understand how Marr's levels force these different dimensions to vary together, though they should be independent in theory (though not always in practice).

Realist Commitments (or, What Does It Mean to Be a Cognitive Realist?)

The first dimension is arguably the easiest to understand: the extent of realism about the theory is simply which parts of the theory are supposed to refer to representations or cognitive processes that "really exist" in a standard metaphysical sense. As a simple example, consider a cognitive model of an individual being asked to add two plus two, and then responding with four. A completely minimal realist commitment for such a model would be to regard it instrumentally: one could commit only to the model offering a correct characterization of the input-output function for human addition. A substantially more realist commitment would claim that there are internal cognitive representations of the numbers '2' and '4', as well as some process by which the former representation (perhaps with a copy) is manipulated so as to yield the latter representation. This interpretation presupposes that there is really a representation there (in a sense discussed below) and that there is some process corresponding to addition.

As we see in this example, simply giving the mathematical specification of a cognitive theory is insufficient to determine the realist commitments; those are, in an important sense, outside of the scope of the computational part of the model. At the same time, to fully understand how to interpret a cognitive model, one needs to know what realist commitments to attribute to it. Such specification rarely occurs explicitly for theories in cognitive science (or at least, rarely in journal papers), but is nonetheless an important step. Some information about realist commitments can be conveyed implicitly through the variables in the model, or by asserting that the theory holds at some level of description. "Levels" of description are, however, much too coarse to convey potentially fine-grained metaphysical commitments, at least in the sense of stating what things there are held to be in the world.

This dimension of variation is still under-specified, as it is not yet clear which epistemological commitments—commitments about what we could come to learn or know—are implied by attributing "reality" to cognitive representations or processes. We can usefully understand epistemological commitments in terms of the predictions they license, as prediction is at the core of many epistemic activities, including control, learning, inference, and even parts of explanation.

By looking at constraints on prediction, we see that there are two different types of realist commitments in the cognitive sciences—realism about processes, and about representations. A rough characterization of the distinction between representations and processes suffices for capturing realist commitments: *representations* are the relatively stable, persistent objects that encode information, and *processes* are dynamic operations involving those objects that can potentially (but need not) change the state of those objects. That is, representations are whatever encodes information stably over some reasonable timescale, and processes are whatever manipulate that information. This high-level characterization covers most of the standard accounts of cognitive representations and processes; even embodied (e.g., Barsalou, 2008) and dynamic systems (e.g., Port & van Gelder, 1995) theories of representation (or its apparent absence) fit this general schema, if we focus on the structure of the theory rather than the language used to describe it.

Given this distinction, *representation realism* implies commitments about the stability of predictions for different types of cognition that use the information encoded in that representation. If the representation "really exists," then the same object is presumably used for (potentially) many purposes, and so predictions in these different contexts should reflect that shared informational basis. For example, realism about the concept 'DOG' implies that behavior in a categorization task involving dogs should be correlated (in various ways) with performance in a feature inference task involving dogs. More generally, representation realism licenses us to use behavior on one task to make predictions about (likely) behavior on different tasks that use the same representations, at least *ceteris paribus*. Importantly, realism about our cognitive representations does not imply that every one is available for every process; it is certainly possible that we have multiple representational stores, some of which are process-specific. But if the same representation is supposed to be available to multiple processes, then representation realism implies a set of epistemological commitments about correlations or stabilities between predictions about the behaviors that the different cognitive processes generate.

Process realism similarly implies epistemological commitments of inter-prediction correlations and stabilities, but rather for the *same* task given *different* inputs, backgrounds, or environmental conditions. That is, if one is committed to the reality of a given cognitive process, then that process should be stable and persistent in its functioning across a range of inputs and conditions. For example, realism about a particular process theory of concept learning implies that this particular process should be active for a variety of inputs that trigger concept learning. Whether I am learning about the concept 'DOG' or the concept 'CAT', the same process should be engaged (since that is the process that is "really there"). Of course, process realism does not imply that every process is triggered for every input or in every condition; rather, process realism is the more minimal claim that there should be correlations and stabilities between the predictions for the different performances of the same task, *ceteris paribus*.

Critically, the epistemological commitments of process realism and representation realism are separable, at least in the abstract. One could think that the appropriate predictive correlations obtain within a cognitive task but not between them (i.e., process realism without representation realism). For example, performance on a categorization task involving dogs might not imply anything stable for predictions about how people do causal inference about dogs. Alternately, the appropriate stabilities might obtain across tasks for the same information, but not within a task (i.e., representation realism without process realism). For example, there might be correlations between predictions for categorization and feature inference tasks involving dogs, but no stable correlations between the predictions for categorization involving dogs and cats.

One can make realist commitments about only some of the representations or processes in one's theory; process and representation realism are not all-or-nothing affairs. To take a concrete example, consider associative models of contingency (or causal) learning, such as the well-known Rescorla-Wagner model (Rescorla & Wagner, 1972). At a high level, associative learning models posit that one learns contingencies or correlations (possibly including causal strengths) by updating associative strengths between various factors. Computationally, whenever one observes a new case, the cognitive agent (i) uses some of the observed factors to predict the state of other factors using the appropriate associative strengths, and then (ii) changes associative strengths based on the prediction error.

Most standard interpretations of associative learning models are realist about the associative strengths, but not about the predictions "generated" in step (i) in order to change strengths in step (ii). That is, the former representations "really exist" and are encoded somewhere, but the latter are just a computational device. Similarly, most are realist about the update process that changes the associative strengths, but not about the prediction process that uses some of the associative strengths to predict the states of other factors.

Degree of Approximation

A second dimension of variation in the commitments of a cognitive theory is in the intended closeness (to reality) of the theory's approximations. All theories are approximate in some ways, in that they exclude certain factors or possibilities; there is no complete theory that incorporates everything. We can nonetheless distinguish (for a particular theory) different commitments about what is *supposed* to be captured by that theory. We can think about this dimension as tracking either which factors have been excluded, or the intended scope of the theory.

As a concrete example, suppose one has a model of human addition that predicts that people will respond '93' when asked "what is 76 + 17?" A question thus arises when someone responds (erroneously) '83': what does this behavior imply for the theory? One response is to hold that this represents a (partial) falsification of the model, as it made a prediction that was not borne out. A different response is to argue that the behavior is due to some factor that was not included in the model because it falls outside of the intended scope of the model (e.g., a momentary lapse of reason due to distraction). The mathematical or computational specification of a theory does not include what was (deliberately) omitted, but that information is important when deciding how to respond to an apparent mismatch between theory and reality.

This dimension is clearly related to the performance/ competence distinction, but it is also not identical with it. Roughly speaking, a competence theory aims to characterize what people are capable of doing, while a performance theory aims to describe what they actually do. Typically, the former is a theory that aims to explain and predict people's

ideal behavior if they did not face, for example, limits on memory and attention, cognitive processing errors, and other deleterious factors. The latter is supposed to be a theory that accounts for these various factors so as to capture (approximately) actual human behavior in all its messy glory. The mathematical specification of a theory does not entail that it is either a performance or competence theory, and some historical debates in the cognitive sciences occurred precisely because of a misunderstanding about whether (the mathematical specification of) a theory was intended as a competence or performance theory.

The performance vs. competence distinction can be understood as picking out two possible commitments along this dimension of variation (i.e., about the intended scope of a theory). But there are many other intended approximations that one could have in mind, including ones that arise from abstracting away from only some human cognitive limitations and peculiarities, rather than all of them (as in competence theories). The performance vs. competence theory distinction marks an important pair of possible intended commitments of a theory, but fails to capture the full range of possible commitments.

Importance of Optimality

The third dimension of variation in a theory's intended commitments is in the putative or claimed optimality of the theory (if any): that is, is the theory additionally claimed to be optimal (or rational), and if so, for what task(s) and relative to what competitors? This additional claim is important because claims about optimality (help to) license so-called "why-explanations." We are often interested not just in *how* some behavior occurs (i.e., the underlying representations and processes that actually generate it), but also in *why* that behavior occurs.

Actually tracing the causal history (whether ontogenetic or phylogenetic) of a process or representation can be remarkably difficult, if not impossible. An alternative path to reach a why-explanation is to show that some cognition is optimal relative to competitors, and that there are sufficiently strong pressures on the individual (or lineage) to push the individual to the optimal cognition (and that those pressures actually obtained in these circumstances). If these elements can be shown, then we can conclude that the cognition occurs because it is optimal. This alternative path is a standard way to demonstrate, for example, that some physical trait constitutes an evolutionary adaptation (Rose & Lauder, 1996).

In practice, many optimality-based "explanations" in the cognitive sciences fail to demonstrate all of the elements; in particular, they frequently fail to show that there are actual "selection pressures" that would suffice to drive an individual towards the optimal cognition, or even to maintain an individual at the optimal cognition. Nonetheless, the intended closeness to optimality (relative to a class of alternatives) of a theory—and so its ability to function in a possible why-explanation—is a critical theoretical commitment about a model that is not implied simply by its mathematical/computational specification. And clearly, variation in this dimension induces different metaphysical and epistemological commitments, as claims that some theory is optimal imply facts about the causal history of the cognition, and about how the cognition should plausibly change under variations in the environment or learning history.

Connecting the Dimensions and Marr's Levels

Marr's levels force these three dimensions of variation to change together, rather than allowing them to vary independently. For example, a theory at the computational level is understood to have a relatively weak set of realist commitments (particularly about processes), significant approximation (since the theory is about how the system should solve a problem, rather than what it actually does), and a fairly strong expectation of optimality. Theories at the implementational level, in contrast, are strongly realist (since they hopefully focus on the underlying biological mechanisms), aim to minimize approximation by incorporating relatively contingent influences, and emphasize causal mechanisms ("how") rather than optimality ("why").

As a result, one must be careful about using Marr's levels to characterize a theory. Use of the terminology can force proponents of a theory into particular commitments that they would prefer to deny, as the levels bundle together commitments that should be kept separate. At the same time, anything that encourages more precise specification of the extra-computational commitments for a theory is a positive. The overall usefulness of Marr's levels principally depends on whether the theory's proponent happens to endorse one of the limited sets of possible commitments that can be expressed in that trichotomy. In many actual cases in cognitive science, however, we have subtler, more fine-grained variations in our theoretical commitments.

From Reduction to Constraint

Throughout this discussion, I have largely ignored one of the most important uses of levels, whether Marr or otherwise: namely, they provide a framework in which we can understand *inter*theoretic relationships. That is, we care not only about the commitments of a scientific theory, but also about the ways in which theories are related to one another, and "levels talk" provides an excellent way to understand such relations.

Of course, it is possible that there are no such (interesting) intertheoretic relations in cognitive science, as implied by various claims that psychology is "autonomous" (or other related term) from the underlying neuroscience (e.g., Fodor, 1974, 1997). Proponents of rational analyses often suggest a similar sort of disconnect, as they sometimes hold that the rational analysis says *nothing* about how the behavior is generated (e.g., Anderson, 1990). There are many theoretical concerns about the autonomy position (see, e.g., the long list in Bickle, 1998). In addition, it is arguably descriptively incorrect: cognitive scientists frequently attend

to the ways in which their theories matter for one another. Regardless of whether it is logically necessary that there be interesting intertheoretic relations, it certainly seems to be contingently true that there are such relations.

The more common way to think about intertheoretic relations in cognitive science is in terms of reduction: roughly, a theory *H* at a higher level must (eventually, somehow) "reduce" to a theory *L* at a lower level. More precisely, *H* reduces to *L* when the latter is a finer-grained version of (something approximately equivalent to) the former. There are many different ways of explicating "reduction" with more precision, whether in terms of syntactic equivalence (Nagel, 1961); semantic equivalence (Bickle, 1998); similar causal powers (Schaffner, 1967); replaceability (Churchland, 1985; Hooker, 1981a, 1981b); or even as implementation of a computer program (Danks, 2008). In all of these cases, there is a close connection, or at least sympathy, between talk of "levels" and the focus on reduction as the key intertheoretic relation.

At least two general concerns arise, however, for all of these accounts of "reduction." First, scientific practice (particularly in the cognitive sciences) often does not involve definite, positive, theoretical proposals to serve as the relata of the "reduction" relation. One might claim, for example, that two variables are associated, or that some functional relationship falls in some (perhaps large) family, or that some previously considered theoretical possibility is incorrect (but without any further information about which theoretical possibility actually is right). These different types of theoretical claims can all imply commitments at other levels even if there is no particular broad theory in which they fit (and so no appropriate relata for reduction).

Second, and more importantly, "reduction" is always understood as a *between*-level relation: *H* and *L* are theories at different levels about roughly similar phenomena.[1] Intertheoretic relations arise, however, between theories that do not stand in this type of "hierarchical" arrangement. For example, theories of causal learning and reasoning (e.g., Cheng, 1997; Griffiths & Tenenbaum, 2005) and theories of "causal" concepts (e.g., Rehder, 2003a, 2003b) investigate different phenomena, and so cannot possibly stand in a reductive relationship in either direction. Nonetheless, these types of theories clearly constrain one another; at the very least, they both depend on representations of causal structure, and so information about one theory can be informative about the other. The focus on "levels of description" or Marr's levels makes it easy to focus on the hierarchically structured theories, but they are not the only ones that constrain one another. Just as we needed a more sophisticated understanding of the dimensions of variation in theoretical commitment, we need a more general account of intertheoretic constraints.

[1] We also sometimes speak of a more general theory "reducing" to a more specific one at the same level in particular conditions (e.g., general relativity reduces to Newtonian dynamics in the limit of $(v / c)^2 \rightarrow 0$). Nickles (1973) shows how to keep this type of reduction separate from the type I have been discussing.

Towards an Account of "Constraint"

At a high level, one cognitive theory *S* constrains another theory *T* if the extent to which *S* has some theoretical virtue *V* (e.g., truth, predictive accuracy, explanatory power) is relevant for the extent to which *T* has the same theoretical virtue *V*. More colloquially, *S* constrains *T* just when, if we care about *T* along some dimension, then we should also care about *S* along that same dimension (because *S* could be informative about *T*). Suppose, for example, that *T* reduces to *S*. Reductions clearly involve constraint in terms of truth: *S* and *T* plausibly have the same truth-value when *T* reduces to *S*. At the same time, reductions arguably do not always involve constraint in terms of explanatory power: the explanatory powers of the two theories in a reduction can vary relatively independently. Thus, it is important to relativize each application of intertheoretic constraint to a particular theoretical virtue.

To see how a more general notion of "constraint" could be made precise, consider the theoretical virtue of truth. I propose (without argument) that: *S* truth-constrains *T* if and only if a change in belief in *S* from time t_1 to time t_2 would, for a fully-knowledgeable agent, *rationally* produce a change in belief in *T* from t_1 to t_2. Note that there is no assumption here that the change in belief in *S* is rational; rather, this account of 'constraint' essentially models it as a conditional: "if an individual's belief in *S* changes (for whatever reason), then belief in *T* should rationally change as well, assuming that she understands the implications of her beliefs."

This proposal clearly includes reduction as a special case constraint: if *H* reduces to *L* given conditions *C*, then an increase in belief in *L&C* (alternately, full acceptance of *L&C*) should rationally lead to an increase in belief in (or full acceptance of) *H*. For example, if some psychological theory *P* reduces to some neuroscientific theory *N*, then if we come to believe *N*, then we should also (rationally) believe *P*. Moreover, in some contexts, a reductive relation can also lead to a downward constraint: if we come to believe *H*, then that can rule out certain *L*s (i.e., any that *H* cannot reduce to).

This account of truth-constraint applies much more broadly than just reduction. For example, causal learning theories and theories of causal concepts that use the same representational framework (e.g., causal Bayesian networks) can be understood as mutually supporting: each makes the other more probable. More generally, one regularly finds arguments in cognitive science that are based on converging evidence from disparate domains, measurement methods, or processes. In this model of truth-constraints, the theories in the different domains place symmetric constraints on one another: increases (or decreases) in belief in one theory should rationally lead to increases (or decreases) in belief in others that point in the same direction. That is, the broader intertheoretic relation of "constraint" enables us—in contrast to the more narrowly focused "reduction"—to explicate and justify one of the most common argumentative techniques in cognitive science.

Conclusions

The core idea of this paper is that the commitments that we have about our cognitive theories extend far beyond their mathematical or computational specification. Instead, we must be clear about where we are located in a multi-dimensional space of theoretical commitments. Our degree of realist commitment, permissible degree of approximation, and intended degree of optimality all can vary relatively independently, though they are tightly coupled in the traditional Marr levels.

Moreover, we need a more fine-grained notion of intertheoretic relations to complement this more nuanced picture of theoretic commitments. Cognitive theories sometimes reduce to one another, but more commonly they inform one another only indirectly. I have suggested that a theory of intertheoretic constraints would be most appropriate, but have only sketched how such constraints might look in one particular case. Substantial work remains to be done to characterize the ways that theories can relate to one another, and then to show how these constraints can be used to guide actual practice in cognitive science.

Acknowledgments

This work was partially supported by a James S. McDonnell Foundation Scholar Award. An early version of this paper was presented at the 2013 Marshall Weinberg Cognitive Science Symposium.

References

Anderson, J. R. (1990). *The adaptive character of thought*. Hillsdale, NJ: Erlbaum.

Barsalou, L. W. (2008). Grounded cognition. *Annual Review of Psychology, 59*, 617-645.

Bechtel, W. (1994). Levels of description and explanation in cognitive science. *Minds & Machines, 4*, 1-25.

Bickle, J. (1998). *Psychoneural reduction: The new wave*. Cambridge, MA: The MIT Press.

Cheng, P. W. (1997). From covariation to causation: A causal power theory. *Psychological Review, 104*, 367-405.

Churchland, P. M. (1985). Reduction, qualia, and the direct introspection of brain states. *Journal of Philosophy, 82*, 1-22.

Danks, D. (2008). Rational analyses, instrumentalism, and implementations. In N. Chater & M. Oaksford (Eds.), *The probabilistic mind: Prospects for Bayesian cognitive science*. Oxford: Oxford University Press.

Danks, D., Griffiths, T. L., & Tenenbaum, J. B. (2003). Dynamical causal learning. In S. Becker, S. Thrun, & K. Obermayer (Eds.), *Advances in neural information processing systems 15*. Cambridge, MA: MIT Press.

Fletcher, P. C., Anderson, J. M., Shanks, D. R., Honey, R. A. E., Carpenter, T. A., Donovan, T., Papadakis, N. & Bullmore, E. T. (2001). Responses of human frontal cortex to surprising events are predicted by formal associative learning theory. *Nature Neuroscience, 4*, 1043-1048.

Fodor, J. A. (1974). Special sciences: Or the disunity of science as a working hypothesis. *Synthese, 28*, 97-115.

Fodor, J. A. (1997). Special sciences: Still autonomous after all these years. *Nous, 31*, 149-163.

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

Hooker, C. A. (1981a). Towards a general theory of reduction, part I: Historical and scientific setting. *Dialogue, 20*, 38-59.

Hooker, C. A. (1981b). Towards a general theory of reduction, part II: Identity in reduction. *Dialogue, 20*, 201-236.

Marr, D. (1982). *Vision*. San Francisco: W.H. Freeman.

McClamrock, R. (1991). Marr's three levels: A re-evaluation. *Minds & Machines, 1*, 185-196.

Nagel, E. (1961). *The structure of science: Problems in the logic of scientific explanation*. New York: Harcourt.

Nickles, T. (1973). Two concepts of intertheoretic reduction. *The Journal of Philosophy, 70*, 181-201.

Port, R. F., & van Gelder, T. (1995). *Mind as motion: Explorations in the dynamics of cognition*. Cambridge, MA: The MIT Press.

Rehder, B. (2003a). A causal-model theory of conceptual representation and categorization. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 29*, 1141-1159.

Rehder, B. (2003b). Categorization as causal reasoning. *Cognitive Science, 27*, 709-748.

Rescorla, R. A., & Wagner, A. R. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In A. H. Black & W. F. Prokasy (Eds.), *Classical conditioning II: Current research and theory*. New York: Appleton-Century-Crofts.

Rose, M. R., & Lauder, G. V. (Eds.). (1996). *Adaptation*. San Diego: Academic Press.

Satpute, A. B., Fenker, D. B., Waldmann, M. R., Tabibnia, G., Holyoak, K. J. & Lieberman, M. D. (2005). An fMRI study of causal judgments. *European Journal of Neuroscience, 22*, 1233-1238.

Schaffner, K. (1967). Approaches to reduction. *Philosophy of Science, 34*, 137-147.