Evaluating Theories[1]

Simon Dennis

University of Adelaide

Walter Kintsch

University of Colorado

All theories are false (Popper 1959). So in one sense evaluating theories is a straightforward matter. However, some theories are more false than others. Furthermore, some theories have characteristics that tend to promote the advance of scientific knowledge. In this chapter, we will examine what some of those characteristics are and how one goes about the process of identifying and building useful theories.

A theory is a concise statement about how we believe the world to be. Theories organize observations of the world and allow researchers to make predictions about what will happen in the future under certain conditions. Science is about the testing of theories and the data that we collect as scientists should either implicitly or explicitly bear on theory.

There is, however, a great difference between theories in the hard sciences and theories in the soft sciences in their formal rigor. Formal theories are well established and incredibly successful in physics, but they play a lesser role in biology, and even less in psychology where theories are often stated in verbal form. This has certainly been true historically, but some scientists, especially physicists, as well as laypeople, construe this fact to mean that formal theories are restricted to the hard sciences, particularly physics, while formalization is unattainable in the soft sciences. There is absolutely no reason to think so. Indeed, this is a pernicious idea that would permanently relegate psychology to

[1] We would like to thank Michael Lee and Dan Navarro for for many valuable suggestions.

second-class status. If nature is governed by causal laws, they govern both the simplest physical phenomena (which, when analyzed at the level of quantum mechanics turn out not to be so simple at all) as well as the higher-order phenomena in biology and psychology. Human behavior is neither vague nor indeterminate, and hence should be described with equal rigor as physical phenomena.

While psychology is dominated by verbal theories, formal mathematical and computational models have become increasingly popular. Hintzman (1991) argues that mathematics “works” in psychology and many other scientific domains because it enforces precision and consistency of reasoning which for many reasons (such as working memory restrictions or hindsight biases) is not otherwise guaranteed. In recent years, there has been a great deal of interest in not only how one might formulate models mathematically, but how one can formally test theories. This development is particularly important in an area such as psychology where data are inevitably influenced by many sources of noise, so that model selection is a nontrivial exercise (Pitt, Myung TICS). In this chapter, we will give some of the basic ideas behind these developments and provide pointers for those who are interested in understanding these methods in depth.

At the outset we would like to distinguish between the characteristics that lead a theory to be successful from those that make it truly useful. While these are not completely disjoint sets, the literature on the philosophy and sociology of science contains many demonstrations of how factors such as social aptitude, rhetorical power, scientific networks and the sheer unwillingness of proponents to die have led one or another framework to be favored for a time, when in hindsight that has seemed unwise (Kuhn 1962, Gilbert & Mulkay 1984). In this chapter, however, we will focus on those considerations that there is some consensus *ought* to play a role in theory evaluation.

These characteristics include:

1. *Descriptive adequacy.* Does the theory accord with the available behavioral, physiological, neuroscientific and other empirical data?
2. *Precision and interpretability*: Is the theory described in a sufficiently precise fashion that other theorists can interpret it easily and unambiguously?
3. *Coherence and Consistency*: Are there logical flaws in the theory? Does each component of the theory seem to fit with the others into a coherent whole? Is it consistent with theory in other domains (e.g. the laws of physics)?
4. *Prediction and Falsifiability*: Is the theory formulated in such a way that critical tests can be conducted which could reasonably lead to the rejection of the theory?
5. *Postdiction and Explanation*: Does the theory provide a genuine explanation of existing results?
6. *Parsimony*: Is the theory as simple as possible?
7. *Originality*: Is the theory new or is it essentially a restatement of an existing theory?
8. *Breadth*: Does the theory apply to a broad range of phenomena or is it restricted to a limited domain?
9. *Usability*: Does the theory have applied implications?
10. *Rationality*: Does the theory make claims about the architecture of mind that seem reasonable in light of the environmental contingencies that have shaped our evolutionary history?

In different areas of psychology these criteria apply to different degrees and in different ways and it is important to be familiar with the standards in your area. In the following sections, we will give an overview of each criterion and illustrate it with examples. These examples will be drawn primarily from memory and comprehension research, which are our areas of expertise. Similar considerations, however, apply across the discipline.

Criteria on which to Evaluate Theories

Descriptive adequacy: The first and probably most important criterion on which to judge a theory is the extent to which it accords with data. Across psychology data takes many forms. Traditionally, data has been generated through surveys, laboratory experimentation or physiological measures. In areas such as discourse studies and corpus psycholinguistics, however, data come primarily in the form of texts or transcripts of conversations. Furthermore, neuroanatomical studies and brain imagining (including event related potentials, positron emission tomography and functional magnetic resonance imaging) are playing an increasingly important role in psychological theorizing. What is the “right” type of data is a contentious issue and different domains of psychology have different rationales for the way they employ data. Across domains, however, the importance of data is realized and theories that are consistent with the known data are to be preferred.

In psychology, the most popular way of comparing a theory against data is null hypothesis significance testing. Hypothesis testing involves generating two competing hypotheses, one of which would be true if the theory is correct and one which would be false. For instance, our theory of recognition memory might suggest that if we increase the number of items in a study list we will see a decrease in the performance at test. We might then design an experiment in which we present study lists of different lengths to subjects and then ask them to determine which items from a test list appeared on the study list. Using methods of null hypothesis significance testing we can then decide whether the data we collect supports the conclusion that there is a difference or not. In this way, we have tested the theory against data.

The case of list length in recognition memory is an interesting one because it demonstrates that determining whether a theory is consistent with data is not always as straightforward as it may at first appear. It has long been assumed that a decrease in performance was indeed a necessary result of increasing the number of items on the study list. Most existing models predict that length would have this effect and several studies seemed to suggest that these predictions are confirmed (Gronlund & Elam 1994). However, Dennis and Humphreys (2001) proposed a model which did not show a decrement in performance as a consequence of list length. Rather they proposed that variables that are often confounded with length such as the time between study and test and differences in the distribution of attention between short and long lists were responsible for previous results. They conducted experiments that controlled the confounding variables and failed to find any difference between the short and long lists.

This result is controversial (Cary & Reder 2003) and it will likely be sometime before consensus is reached on the actual state of affairs. However, this episode illustrates some of the subtleties involved in determining the extent to which a theory accords with data. One reason that this result continues to be questioned is that using null hypothesis significance testing it is not possible to conclude that there is no difference. A proponent of a theory that predicts a list length effect can always propose that a failure to find the difference was a consequence of lack of power of the experimental design. Perhaps there were not enough subjects to reliably find a result, or perhaps the effect is there but it is small. And the debate remains unresolved.

Another difficulty with null hypothesis significance testing is that it encourages a game of twenty questions with nature (Newell 1973). A study proceeds by setting up one of more dichotomies and at the end we have a set of yes/no answers to these questions. The danger is that rather than develop a cohesive theory and use hypothesis testing to evaluate it, researchers will generate an endless set of issues each of which is only loosely coupled with previous work and little cumulative progress will be achieved (Newell 1973). Not everyone agrees that this has actually been a problem for psychology, but Newell certainly makes a strong case for his argument.

The dominant role of null hypothesis significance testing in psychological investigation has come under intense scrutiny in recent years (Cohen 1994), and in 1999 the American Psychological Association (APA) published recommendations about how one should conduct statistical analyses that included reducing reliance on the hypothesis test (Wilkinson et. al. 1999). Anyone intending to conduct psychological research ought to be familiar with the principles set out in this report.

One of the advantages of formal models of psychological phenomena is that they can be used to derive measures of how well a theory fits the data (Pitt & Myung 2002) that do not rely on null hypothesis significance tests. Typically, a model defines one or more parameters, which are chosen so that the predictions of the model are as close as possible to the observed data[2]. The advantage of the formal model is that we can say exactly how closely it approximates the data (although see the section of parsimony for a discussion of the difficulties with relying on fit alone). In addition, rather than acquiring a set of yes/no answers using formal models gives us additional information about the nature of the relationship between variables as given by the values of the parameters.

[2] Or in some cases the parameters are varied according to some prior distribution to define a distribution of predicted outcomes.

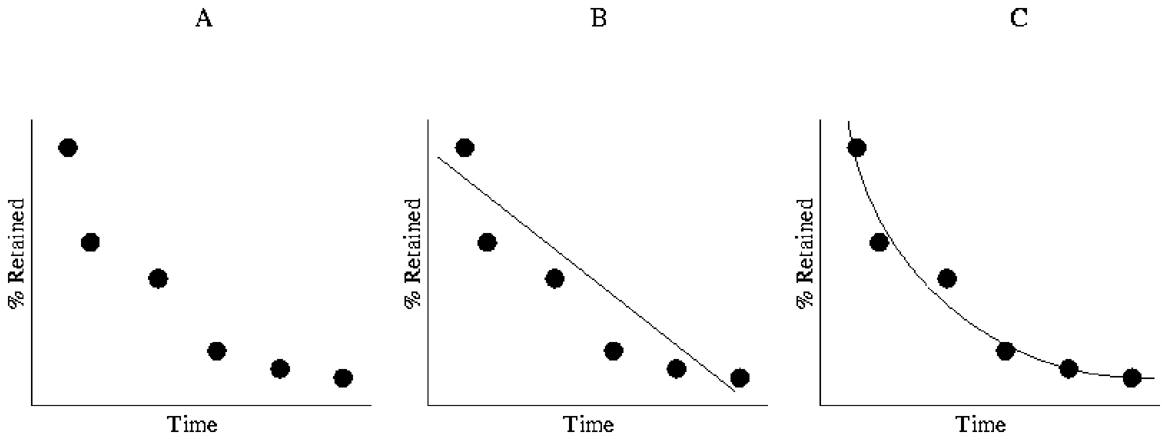


Figure 1: Hypothetical retention data and fits to two possible models. A) Data, B) linear fit C) power fit.

To see how such an investigation might proceed consider how information is retained as a function of time. Figure 1A shows hypothetical data that would be collected from a memory test in which subjects are asked to recall a list of items at different intervals post study. Performance is measured by the number of items they are able to recall. As time progresses we would expect performance to decrease as indicated in the figure. If we were to employ hypothesis testing we could take two of these intervals and ask whether performance decreased.

Using formal models, however, we can ask additional questions. Figures 1B and 1C show two alternative models of how the retention interval and performance are related. In Figure 1B performance decreases as a linear function of the interval and in Figure 1C performance decreases as a power function. So, now we can ask which of these models better describes the form of the relationship between the variables. In this case, we would choose the power function as it is, in general, closer to the data. In addition, we can look at the values of the parameters of the best fitting models to provide an indication of how much we expect performance to decrease across any given interval – that is we can make quantitative predictions. We now have a much deeper understanding than was afforded by the null hypothesis significance test. In particular, we are able in a much more straightforward way to say what it means for the theory to accord with the data.

Precision and interpretability: One common criticism of psychological theories is that they are often described in vague and imprecise language making them difficult for independent researchers to apply as the following quotes attest:

“Terms are used such as “access to the lexicon”, “automatic processing”, “central executive”, “resources”; formal definitions of such terms are rare, and even rarer are statements of the rules supposed to be governing their interaction. As a result one is left unclear about exactly what kinds of experimental data would invalidate such theories, and whether or not they are intended to apply to some new experimental situation” (Broadbent, 1987, p. 169) and

“When a theory does attract criticism, the critic almost always turns out to have misunderstood, and the theory stands as originally proposed... On the rare occasion a criticism demands action, fine tuning will almost always suffice. Thus, the chance of a theory having to be abandoned or even appreciably revised as a consequence of criticism are vanishingly small, and hence researchers can be confident that their theories will stay alive just as long as they continue to nourish them (p. 328, Watkins, 1990).

When evaluating a theory pay close attention to how well the constructs in the theory and the relationships between them are defined. Can you be confident that you could apply the theory in a related domain unambiguously? Conversely, when articulating a theory ask yourself what implicit assumptions you may be making that may not be shared by your readers.

Coherence and Consistency: Another hallmark of a good theory is its coherence and consistency. While it may seem obvious that theories should be free from logical flaws it isn't always easy to spot these flaws particularly when theories are presented in verbal form.

As an example (from Hintzman 1991), consider the following claim taken from a textbook on learning "While adultery rates for men and women may be equalizing, men still have more partners than women do, and they are more likely to have one night stands." (Leahey & Harris, 1985, p. 287). On the surface the statement seems plausible and consistent with our understandings of male and female behavior. However, a more careful consideration reveals a logical flaw. From context it was clear that the claim did not hinge on homosexual encounters, nor did it rely on small differences in the overall numbers of males and females, which would have rendered it uninteresting. So, as Hintzman points out, given that each liaison will usually involve one man and one woman, there really can't be a substantive difference in either the number of partners or the number of one night stands between the genders. Of course there may be other differences, such as the willingness to report, but there can't be an actual difference in the total counts for each of the genders.

Another common problem that one must look for in evaluating theories is circularity and again the issue can be quite subtle. One such example is the notion of Transfer Appropriate Processing (TAP) that appears in the memory literature. TAP asserts that "performance is a positive function of the degree of overlap between encoding and retrieval processes" (Brown & Craik 2000, p 99). So, the more similar the processes evoked at test are to the processes evoked during encoding of the material the more likely it is that information will be retrieved. Again on the face of it this seems like a reasonable theoretical conjecture. However, the difficulty arises in defining what it means to have overlap in unseen psychological processes. If the only mechanism for determining overlap is performance itself, then, the overlap in processes determines performance, but performance is our benchmark for determining if there is an overlap in processes. To the extent that this is actually the case, the TAP claim becomes vacuous.

Beyond ensuring that theories are free from logical flaws of the kind illustrated above, it is also important to ask how consistent a theory is both with other theories within psychology and also with theory outside psychology. For instance, one might prefer a theory of text comprehension that incorporates the constraints on working memory capacity that have been found by memory researchers (Kintsch 1998). For the same reason, theories of psi and remote viewing are dispreferred because they are inconsistent with physical laws. Of course, it is always possible that our understanding of working memory or even the physical universe will change in ways that would invalidate our theoretical assumptions. However, our current understanding remains our best guess at how the world operates and so theories that are consistent with this approximation are more likely to endure.

Prediction and Falsifiability: One of the key attributes of a good theory is falsifiability (Popper 1959). Ideally, one should be able to make unambiguous predictions based on the theory and conduct empirical tests that could potentially bring the theory into doubt. We start this section by giving an example of falsification in action.

Many models of recognition memory, particularly those known as global matching models (SAM, Gillund & Shiffrin 1984, TODAM, Murdock 1982, Minerva II, Hintzman 1984, Matrix model, Humphreys, Bain & Pike 1989, see also Humphreys, Bain, Pike & Tehan 1989) propose that performance is compromised by other items that appear in the same context. Ratcliff, Clarke and Shiffrin (1990) realized that one of the consequences of this assumption is that as the strength of one item on a study list is increased either by increasing the time for which it is studied or by increasing the number of times it appears performance should decrease for the other items in the list. Because the global matching models are mathematical models it is possible to formally prove this prediction.

In series of experiments, however, Ratcliff et. al. (1990) demonstrated that no such effect occurs. They presented subjects with a list consisting of items each of which was presented once or alternatively with a list in which some items appeared once and some items were repeated. The items that are repeated show improved performance, but there was no difference between the once presented items as a function of the strength of the other items. This result immediately falsified this entire class of models and has led to a productive time in the area as researchers look for alternative models that are capable of accounting for this data.

While falsification provides the most useful information in advancing scientific knowledge, it is sometimes the case that verifying predictions can increase our confidence in a theory. However, not all predictions are equally diagnostic. Predictions that seem to violate our intuitions and yet turn out to be the case provide more support for a theory than predictions which are unsurprising.

A classic example of such a prediction is the phenomenon of over-expectation in fear conditioning. The Rescorla-Wagner model of classical conditioning (Rescorla, 1970) proposes that the amount of learning that occurs on a trial when a tone is paired with a shock is proportional to the difference between the maximum value of associative strength and the current strength of the association between the tone and the shock. This simple model provides a good account of a number of conditioning phenomena including how responding increases as a function of learning trials, how responding decreases if the tone is presented without the shock (extinction), and the lack of learning when a novel conditional stimulus is presented with another conditional stimulus that has already been associated with the unconditional stimulus (blocking).

Being able to account for these phenomena with such a simple model was impressive in its own right. However, Rescorla (1970) went one step further: he deduced a highly non-intuitive prediction from his model and tested it experimentally. He reasoned that if two stimuli that each by itself was highly associated with the unconditioned stimulus were jointly presented, then the response to this compound cue should decrease in strength because the difference between the total associative strength of the two cues and the maximum strength would be negative. This is a startling prediction because intuition would suggest that any pairing of the conditional and unconditional stimuli should produce positive learning. However, the model predicted otherwise and Rescorla's (1970) results confirmed the prediction.

Postdiction and Explanation: In general, clear demonstrations of prediction and falsification like those above are rare in psychology. “Predictions are hard to make, especially about the future,” said the Nobel Prize winner Niels Bohr about physics, and that is even truer of psychology. Prediction is possible under well-controlled laboratory conditions, like those outlined above, but hardly ever under natural conditions. We can predict the responses of subjects in the lab, but not where it really counts, in real life. That does not mean that our theories are no good, only that they do not afford prediction. Our explanations of behavior often are not predictive, but only *postdictive* (the term retrodictive is also used). Admittedly, postdictive explanations are weaker than predictive explanations, but they are still explanations. Other sciences, too, rely primarily on postdiction. A good example is meteorology. Meteorologists do predict the weather, but with limited success. That is not because their explanations or theories are not good: it is probably not an exaggeration to say that they understand very well the physical laws that govern the weather, from the molecular level to the macro-level. Yet they are often unable to predict what will happen because to do so perfectly one would have to know the state of the atmosphere with respect to a multitude of variables in exquisite detail and over a huge geographic area. Even knowing what every molecule is doing where, and having a big enough computer to crank out predictions might still not be enough because of chaos effects: a tiny disturbance somewhere – the flutter of a butterfly's wings in Mongolia – could under certain circumstances have system-wide consequences – resulting in a hurricane in Florida. Prediction is hard, indeed, but after the fact the meteorologist can explain quite well what happened and why.

Psychology is in a similar situation: there no reason to think that we shall ever have available all the information about a person's history and current state that would enable a precise prediction of his future acts. Thus, prediction cannot be our goal, except in limited circumstances. But explanation after the fact – postdiction – is both possible and worthwhile. Much the same can be said about linguistics: linguists have good and formal theories to explain, for instance, phonological change, but the explanation is after the fact. Medicine is another discipline that is basically not predictive, but postdictive: just how and when a particular person will die is in general not predictable, but the good doctors can very well explain what happened and why once he is dead. To predict we need to understand what is going on and have sufficient control over the relevant variables; postdiction also implies understanding, though not control. The true goal of science is understanding, not necessarily prediction.

Postdiction can be based on formal theories as much as prediction. Thus, the difference between the hard and soft sciences is not in the degree of formalization and rigor. For psychology formal rigor is a worthwhile and achievable goal, even though prediction is not generally feasible.

Parsimony (Occam's Razor): The principle of parsimony states that theories should be as simple as possible. This idea was first articulated by the Franciscan friar, William of Ockham as *pluralitas non est ponenda sine neccesitate*, which translates as "Plurality should not be posited without necessity". That is, only those causes or constructs that are needed to explain a phenomenon should be posited.

In the discussion of descriptive adequacy above, we stated that it is important that a model of a theory be able to fit the existing empirical data. While this is certainly a desirable quality it is important to realize that the ability to fit data by itself does not necessarily add a great deal of credibility to a model. It is also important to consider the range of datasets the theory can fit. Some theories are very flexible and are able to fit both data that is observed and data that is not observed. In this case, the ability to fit the data provides little comfort (Roberts & Pashler 2000).

To appreciate this point, consider the retention data that we discussed above. Figure 2 shows this data fit with linear and power functions as above (panels A and B) as well as with a more complicated cubic function (panel C).

In this case, the fit becomes progressively better as we move from linear to power to cubic functions. In general this will be the case because the cubic is a more flexible function and can look similar to the linear and power functions, particularly over a restricted range. It is able to model not only the true underlying function but also the random noise that is part of this dataset but would not appear if we were to run the experiment again.

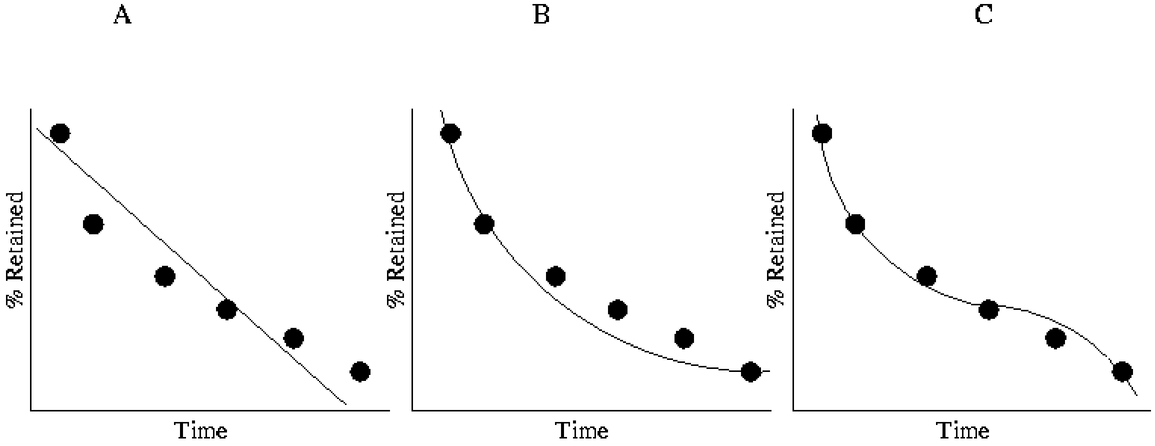


Figure 2: Fits to retention data. A) Linear B) Power C) Cubic. Adapted from lecture slides prepared by Michael Lee.

Consequently, when generalizing to new data the cubic function often does not do well. Figure 3 shows what can happen when we extend the time scale over which we collect data. In the new regions, the power function does a much better job.

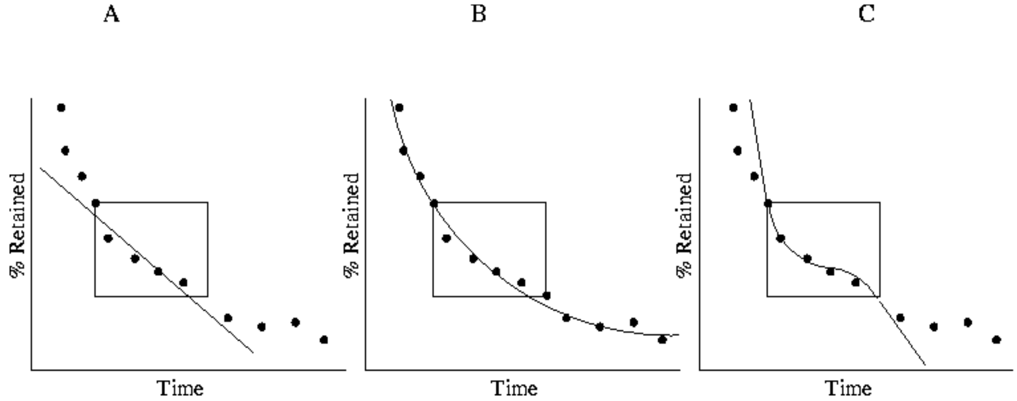


Figure 3: Generalizing to new data. A) Linear B) Power C) Cubic. Adapted from lecture slides prepared by Michael Lee.

So in choosing between models of a given phenomenon, we would like to favor the one that fits the data well AND is as simple as possible. That is, we would like to fulfill the aims of descriptive adequacy and parsimony at the same time.

A number of techniques have been developed to achieve these objectives (see Pitt, Myung & Zhang 2002 and Pitt & Myung 2002 for summaries). They include the Akaike information criterion (AIC; Akaike, 1973), the Bayesian information criterion (BIC; Schwarz, 1978), the information theoretic measure of complexity (ICOMP; Bozdogan, 1990), cross-validation (CV; Stone, 1974), minimum description length (Pitt, Myung & Zhang 2002), Bayesian model selection (BMS; Kass & Raftery, 1995; Myung & Pitt, 1997), the Stochastic Complexity Criterion (SCC: Rissanen 1996), and the Geometric Complexity Criterion (GCC: Myung, Balasubramanian, & Pitt 2000, Navarro & Lee in press). In addition, techniques such as response surface analysis (Myung, Kim & Pitt 2000), landscaping (Navarro, Myung, Pitt & Kim 2003) and the parametric bootstrap cross-fitting method (Wagenmakers, Ratcliff, Gomez & Iverson 2004) provide additional insight into the relationships between models. Outlining these methods is beyond the scope of this chapter, but it is important to be aware that the area of model selection is developing rapidly and is likely to play an increasingly significant role in theory evaluation as it matures.

Breadth: Unlike sciences like physics, in which providing a unified account of phenomena is seen as the ultimate goal, in psychology efforts at unification are often treated with deep skepticism. The modus operandi in psychology tends to be to continually divide phenomena into smaller and smaller classes – to know more and more about less and less. Instead of general theories in psychology, we have miniature models that are expected to hold under only very specific conditions.

Many psychologists believe that this is simply a consequence of the lack of maturity of the science of psychology, which when coupled with the complexity of psychological phenomena, means that we are simply not ready for broad theories that integrate many different phenomena in different areas of psychology. Newell (1973) argued otherwise, however. He felt that the piecemeal approach hampered progress, and that it was time to begin work on a unified theory of psychology. In cognitive psychology, the push for unification has been led by Newell (1990, the SOAR model) and Anderson (Anderson &Lebiere, 1998, the ACT-R model).

Whether one believes that the time for a completely unified theory is upon us, clearly theories should attempt to be as broad as possible, while maintaining the other criteria such as descriptive adequacy and the ability to provide genuine explanations of phenomena.

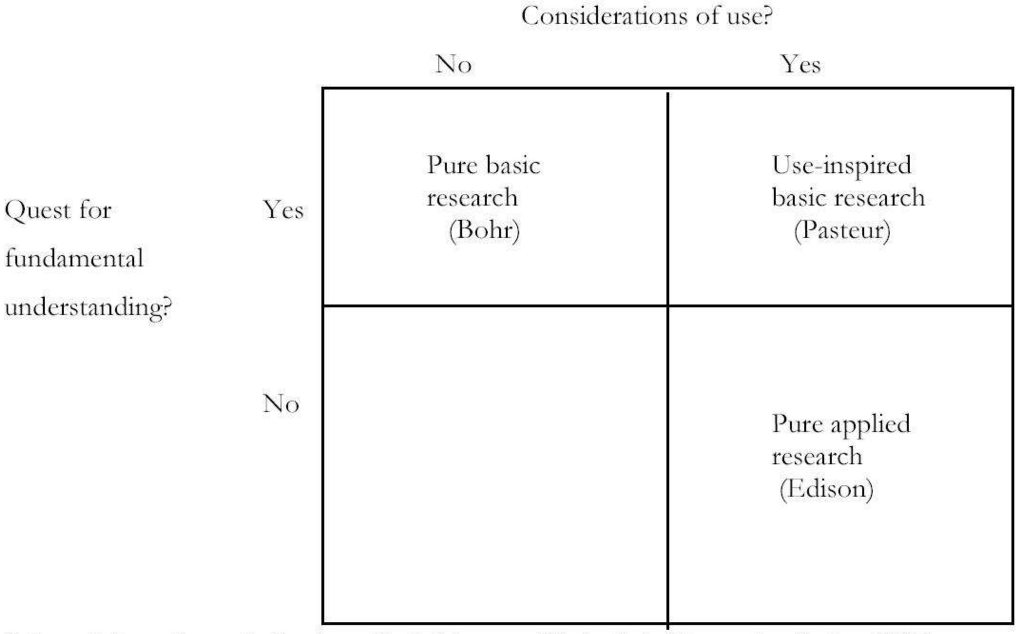
Originality: Another key characteristic of a theory is its originality. It goes without saying that plagiarism is unacceptable. However, originality goes well beyond issues of plagiarism. Theories may look very different, and indeed be different in their broader implications, but yet with respect to a particular set of data it may not be possible to differentiate them. Thus, great care must be taken when comparing theories against each other, even when they are stated formally. A good example of how tricky even straightforward tests between theories can be comes from the early work on concept identification. The experiments were simple: for instance, subjects saw on each learning trial a 5-letter string and had to sort them into two types, A and B. After they made their response, they were told whether they were correct or not. The experimenter employed an arbitrary classification rule, say "If the 4th letter is an R, it is type A; if not it is type B". Researchers already knew that subjects either knew the rule or did not, in which case they guessed. At issue was the question how they learned the rule. According to the old reinforcement view, learning could occur on every trial since the response was reinforced each time. According to the then novel more strategic view of learning, learning could occur only when an error was made, because when subjects were told they were correct, they had no reason to change what they were doing. It seemed an easy matter to be decided. One can construct two simple mathematical models, each with two parameters , a learning rate and a guessing probability, perform an experiment, and fit the proportion correct on each learning trial to both models; the one with the better fit wins. Surprisingly, investigators found out that both models fit equally well. Eventually, it was shown that, in spite of the fact that they are based on opposite assumptions about the nature of learning, the two models made exactly the same predictions as far as the learning data in these experiments was concerned. Their mathematical formulations looked very different, but in fact were just different ways to express the same structure. Fortunately, the story does not end here, but has a happy end: once researchers knew what the underlying problem was, they could design experiments that looked beyond learning curves to data that actually did discriminate between these two views of learning. (It turned out that the learning-or-errors-only model was the correct one for these simple concept learning problems; for a full discussion of this episode in theory construction see Kintsch, 1977, Chapter 7).

Usability: Good scientific theories should be useful in addressing societal problems. In the landmark report "Science: The endless frontier", Vannevar Bush (1945) outlined a linear model of the relationship between scientific discovery and technological innovation which saw basic research and applied research at opposite ends of a continuum. Under this model progress was made by conducting fundamental research without concern for use and then in a quite separate exercise transferring the knowledge gained into technology.

More recently the assumption that research must either be basic or applied has been challenged (Stokes 1997). Figure 4 shows schematically the view that has been emerging in scientific communities and amongst policy makers, namely that considerations of use and the quest for fundamental understanding are different dimensions in the research landscape (Stokes 1997).

Under this view the best research (and the best theory) contributes to scientific understanding while fulfilling a societal need. The work of Luis Pasteur is a prime example of this type of research. Pasteur was a key contributor to the development of the field of microbiology in the 19th century. As Stokes notes, however:

“There is no doubt that Pasteur wanted to understand the process of disease at the most fundamental level as well as the other microbiological processes that he discovered, but he wanted that to deal with silk worms, anthrax in sheep and cattle, cholera in chickens, spoilage in milk, wine and vinegar, and rabies in people.“ (p. 6)



		Considerations of use?	
		No	Yes
Quest for fundamental understanding?	Yes	Pure basic research (Bohr)	Use-inspired basic research (Pasteur)
	No		Pure applied research (Edison)

(adapted from *Pasteur’s Quadrant: Basic Science and Technological Innovation*, Stokes 1997).

Figure 4: A schematic representation of the landscape of scientific research in relation to the search for fundamental understanding and applied significance.

Within psychology there are many examples of working in Pasteur's quadrant. One such example is the theory of meaning called Latent Semantic Analysis (LSA, Landauer & Dumais 1998). LSA is a method for taking large text corpora and deriving vector representations of the words in that corpora based on the patterns of word usage. Words with similar meanings tend to have similar locations in the semantic space defined by the model. From a fundamental perspective, then, the model provides a theory of how meaning is represented and acquired.

In addition, however, LSA has been used extensively in educational technologies. One of the most impressive of these technologies is an automated essay grader. LSA makes it possible to represent texts as vectors, so that an essay that can be compared against a gold standard essay to assess whether a student has captured the essence of a writing assignment. Furthermore, it can do this with a reliability that is equivalent to human tutors (Landauer, Laham & Foltz, 2001). So, the LSA theory is advancing our understanding of human cognition and providing a valuable (and commercially viable) service to the community.

Rationality: The final criteria that we will consider is rationality, that is, does the theory make claims about the architecture of mind that seem reasonable in light of the environmental contingencies that have shaped our evolutionary history? Stated in this way, the criterion seems rather vague. Given enough imagination it is usually possible to envisage evolutionary scenarios to justify most claims about the way the mind should be. However, Anderson (1990) showed that a more precise notion of rationality is possible if one focuses on how information appears in the environment. Anderson argues that the cognitive system is what it is because it is adapted to the way information is distributed in the environment. It is the structure of the environment that over eons of evolution has shaped psychological processes so that they are maximally able to deal with that structure.

As an example, consider the retention curves that we looked at earlier. We concluded at that stage that the power function was the best fit to the data. But why should the memory system be constructed in such a way that performance would degrade as a power function of the retention interval? Anderson and Schooler (1991) showed that the power function is ubiquitous in nature. If you look at the probability that a word or phrase in newspaper articles or child language corpora repeats as a function of the time since its last occurrence, it decreases as a power function of the interval. Forgetting also occurs according to the power law because the memory system has been optimized through evolution to deal with an environment that encapsulates the power law.

It is not always easy to capture the relevant environmental statistics in the way that Anderson and Schooler (1991) did. However, when it is possible it provides compelling support for a theory.

Conclusions

We have outlined in this chapter a number of considerations that should be taken into account in the evaluation of theories. We have also tried to provide access to the literature on evaluating theories for those readers who need to know more. Evaluating theories – like science in general – is hard, however, and there are no recipes for doing so. The issues we have raised here are certainly worth serious consideration, but they are no substitute for critical analysis. Every theory, every model is different and makes different demands on the reader. The points we have discussed here are valid ones in general, but how they are to be weighted in a specific case depends on many factors. Surely, parsimony is a worthwhile criterion for evaluating theories, but equally surely there are cases where the less parsimonious theory should be preferred. For example, if parsimony is bought at the expense of dramatically restricting the scope of a theory, that may not be a good bargain because it conflicts with the goal of having a broadly applicable theory. Thus, a highly parsimonious theory of recognition memory may not be as attractive as a less parsimonious one that applies to all of memory. On the other hand, a theory of everything that relies on a host of free parameters is not very interesting either. Where is the fine line that should not be transgressed? There is none, every case must be analyzed on its own merit and conflicting factors must be weighted carefully anew in each instance. Evaluating theories – and even more so, building theories in the first place – cannot do without creative, original thinking. It has rules and principles, as we outlined here, but they cannot be applied thoughtlessly.

References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In B. N. Petrox & F. Caski (Eds.), Second international symposium on information theory (pp. 267-281). Budapest, Hungary: Akademiai Kaido.

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

Anderson, J. R. & Lebiere, C. (1998). *The atomic components of thought.* Mahwah, NJ: Erlbaum.

Anderson, J. R. & Schooler, L. J. (1991). Reflections of the environment in memory.

Psychological Science, 2, 396-408.

Bozdogan, H. (1990). On the information-based measure of covariance complexity and its application to the evaluation of multivariate linar models. Communications in Statistics: Theory and Methods, 19, 221-278.

Broadbent, D. (1987). Simple models for experimental situations. In P. Morris (Ed.), Modeling Cognition. (pp. 169-185). London: Wiley.

Brown, S. C. & Craik, F. I. M. (2000). Encoding and retrieval of information. In E. Tulving and F. I. M. Craik (Eds.) The Oxford Handbook of Memory (pp. 93-108.), New York: Ny, Oxford University Press.

Bush, V. (1945). Science – The endless frontier, Appendix 3. Report of the committee on science and the public welfare. U.S. Government Printing Office, Washington, D.C.

Cary, M. & Reder, L. M. (2003). A dual-process account of the list-length and strength-based mirror effects in recognition. Journal of Memory and Language, 49, 231-248.

Cohen, J. (1994). The earth is round (*p* < .05). *American Psychologist*, 49, 997–1003.

Dennis, S. & Humphreys, M. S. (2001). A context noise model of episodic word recognition. Psychological Review. 108(2). 452-477.

Gillund, G., & Shiffrin, R. M. (1984). A retrieval model for both recognition and recall. *Psychological Review, 91*, 1-67.

Gronlund, S. D., & Elam, L. E. (1994). List-length effect: Recognition accuracy and variance of underlying distributions. Journal of Experimental Psychology: Learning, Memory and Cognition, 20, 1355-1369.

Hintzman, D. L. (1984). MINERVA2: A simulation model of human memory. *Behavior*

Research Methods, Instruments, and Computers, *16*, 96-101.

Hintzman, D. L. (1991). Why are formal models useful in psychology? In W. E. Hockley and S. Lewandowsky (Eds.) Relating theory and data: Essays on human memory in honor of Bennet B. Murdock. (pp. 39-56). Hillsdale: NJ. Lawrence Erlbaum Associates.

Humphreys, M. S., Bain, J. D. & Pike, R. (1989). Different ways to cue a coherent memory system: A theory for episodic, semantic and procedural tasks. Psychological Review, 96, 208-233.

Kass, R. E. & Raftery, A. E. (1995). Bayes factors. Journal of the American Statistical Association, 90, 773-795.

Kintsch, W. (1977). *Memory and Cognition*. New York: Wiley.

Kintsch, W. (1998). *Comprehension: A paradigm for cognition*. New York: Cambridge University Press.

Kuhn, T. (1962). *The structure of scientific revolutions*. Chicago, IL: University of Chicago Press.

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The Latent Semantic Analysis theory of acquisition, induction and representation of knowledge. *Psychological Review, 104*, 211-240.

Landauer, T. K., Laham, D., & Foltz, P. W. (2000). The Intelligent Essay Assessor. *IEEE Intelligent Systems*, 27-31.

Leahey, T. H., & Harris, R J. (1985). *Human learning*. Englewood Cliffs, NJ: Prentice-Hall.

Murdock, B. B., Jr. (1982). A theory for the storage and retrieval of items and associative information. *Psychological Review*, *89*, 609-626.

Myung, I. J., Balasubramanian, V., & Pitt, M. A. (2000). Counting probability distributions: Differential geometry and model selection. Proceedings of the National Academy of Sciences USA, 97, 11170-11175.

Myung, I. J. & Pitt, M. A. (1997). Applying Occam's razor in modeling cognition: A Bayesian approach. Psychonomic Bulletin & Review, 4, 79-95.

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

Navarro, D. J. & Lee, M. D. (in press). Common and distinctive features in stimulus representation: A modified version of the contrast model. *Psychonomic Bulletin & Review*.

Navarro, D. J., Myung, I. J., Pitt, M. A. & Kim, W. (2003). Global model analysis by landscaping. In *Proceedings of the 25th Annual Conference of the Cognitive Science Society*.

Newell, A. (1973). You can't play twenty questions with nature and win. In W.C. Chase (Ed.). Visual information processing. New York: Academic.

Newell, A. (1990). *Unified Theories of Cognition*. Cambridge, Mass.: Harvard University Press.

Pitt, M. A. & Myung, I. J. (2002). When a good fit can be bad. Trends in cognitive sciences, 6(10), 421-425.

Pitt, M. A., Myung, I. J. & Zhang, S. (2002). Towards a method of selecting amoung computational models of cognition. Psychological Review, 109(3), 472-491.

Popper, K. R. (1959). The logic of scientific discovery. London: Hutchinson.

Gilbert, G. N. & Mulkay (1984). Opening Pandora's Box: A Sociological Analysis of Scientists' Discourse. Cambridge: Cambridge University Press.

Ratcliff, R., Clarke, S., & Shiffrin, R. (1990). The list strength effect: I. Data and discussion. Journal of Experimental Psychology: Learning, memory and cognition, 16, 163-178.

Rescorla, R. A. (1970). Reduction in the effectiveness of reinforcement after prior excitatory conditioning. Learning and Motivation, 1, 372-381.

Rissanen, J. (1996). Fisher information and stochastic complexity. IEEE Transactions ion Information Theory, 42, 40-47.

Roberts, S. & Pashler, H. (2000). How persuasive is a good fit? A comment on theory testing. Psychological Review, 107, 358-367.

Schwarz, G. (1978). Estimating the dimension of a model. The Annals of Staitics, 6, 461-464.

Stokes, D. E. (1997). Pasteur's Quadrant: Basic Science and Technological Innovation, Brookings Institution Press, Washington, DC.

Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions (with discussion). Journal of the Royal Statistical Society Series B, 36, 111-147.

Wagenmakers, E. J., Ratcliff, R. Gomez, P. & Iverson, G. J. (2003). Assessing model mimicry using the parametric bootstrap. *Journal of Mathematical Psychology*, 48, 28-50.

Watkins, M. J. (1990). Mediationism and the obfuscation of memory. American Psychologist, 45, 328-335.

Wilkinson, L. and the Task Force on Statistical Inference APA Board of Scientific Affairs (1999). Statistical Methods in Psychology Journals: Guidelines and Explanations, *American Psychologist*, 54, 594-604.