Continuing Commentary

Commentary on **Philip N. Johnson-Laird & Ruth M. J. Byrne (1993). Précis of *Deduction.* BBS 16:323–380.**

Abstract of the original article: How do people make deductions? The orthodox view in psychology is that they use formal rules of inference like those of a "natural deduction" system. *Deduction* argues that their logical competence depends, not on formal rules, but on mental models. They construct models of the situation described by the premises, using their linguistic knowledge and their general knowledge. They try to formulate a conclusion based on these models that maintains semantic information, that expresses it parsimoniously, and that makes explicit something not directly stated by any premise. They then test the validity of the conclusion by searching for alternative models that might refute the conclusion. The theory also resolves long-standing puzzles about reasoning, including how nonmonotonic reasoning occurs in daily life. The book reports experiments on all the main domains of deduction, including inferences based on propositional connectives such as "if" and "or," inferences based on relations such as "in the same place as," inferences based on quantifiers such as "none," "any," and "only," and metalogical inferences based on assertions about the true and the false. Where the two theories make opposite predictions, the results confirm the model theory and run counter to the formal rule theories. Without exception, all of the experiments corroborate the two main predictions of the model theory: inferences requiring only one model are easier than those requiring multiple models, and erroneous conclusions are usually the result of constructing only one of the possible models of the premises.

Mental model theory and pragmatics

Jean-Baptiste van der Henst
Centre de Recherche en Epistémologie Appliquée, 75005 Paris, France
henst@poly.polytechnique.fr

Abstract: Johnson-Laird & Byrne (1991; 1993) present a theory of human deductive reasoning based on the notion of mental models. Unfortunately, the theory is incomplete. The present commentary argues that pragmatic considerations, particularly of the type discussed in Sperber and Wilson (1995), can complement the theory.

Johnson-Laird & Byrne (1991) (JL&B) conclude their book by claiming that their theory is incomplete (p. 213; all page references are to the book, not the Précis). The present commentary suggests two ways of complementing the theory with some pragmatics. First, contextual factors accounted for by linguistic pragmatics may help predict which conclusion people will derive from the models they construct. Second, a pragmatic approach may help specify how the search for alternative models is triggered.

(1) Even if they do not refer to pragmatics explicitly, JL&B (pp. 21–22) list "extra-logical" constraints governing individuals' conclusions that are actually of a pragmatic nature. However, these constraints do not permit us to make accurate predictions about which conclusion people will try to infer from a set of premises. For example, the authors claims (p. 35) that after constructing a mental model, individuals try to produce "something" not explicitly stated in the premises. The question one can ask is: What kind of "something" do people try to produce? The question becomes relevant, in particular, when one can draw several conclusions from the same set of premises. For example, from the premises

Paul is taller than John,
John is taller than Pete,
Pete is taller that Bob,

one can infer that Paul is taller than Bob, but one can also infer that the four boys are of different heights. Which inference will people make? This is a pragmatic question. The context in which information is processed determines to a great extent which inferences people will make. One can easily assume that individuals will try to draw a conclusion that is the most relevant one in the context of processing (Sperber & Wilson 1995). In the above example, if one is interested in knowing the most important relational contrast, then one will probably infer that Paul is taller than Bob. However, if one wants to know whether some of the boys are of the same height, one will probably infer that the four are of different heights.

JL&B argue (p. 93) that indeterminate premises such as "the circle is in front of the triangle; the cross is behind the circle" do not support a valid conclusion because the two models compatible with the premises do not yield a determinate relation between the triangle and the cross. From these premises, people will probably conclude that "nothing follows." Nevertheless, one logical conclusion is "the circle is in front of the triangle and the cross." This conclusion is compatible with the two models but is of course trivial, and people will probably be reluctant to draw it. But, drawing a conclusion does not only depend on its "nontriviality." It can also depend on the further inferences it may allow.

The premises:
John is ahead of Paul
John is ahead of Bob

support the trivial inference that John is ahead of Paul and Bob. But when the premises are given in the context "*John, Paul, and Bob were the first three finishers of the athletics race last Sunday,*" that inference conveys more cognitive effects (Sperber & Wilson 1987; 1995) than in a neutral context, because it allows one to know that John won the race. The experimental results of a production task (van der Henst et al. 2000) show that in the neutral context, participants give significantly more indeterminate answers (31.4%) such as "nothing follows" than they do in the "race" context (11.1%).

(2) Relevance considerations can also shed light on the procedure of searching for alternative models. According to JL&B, reasoners draw a putative conclusion from an initial model and then try to construct alternative models of the premises. The search for alternative models is the genuinely deductive stage according to the authors (pp. 36 and 127). JL&B seem to link good reasoning with the procedure of searching exhaustively for counterexamples. Reasoners following this procedure are designated as "prudent" (p. 35). Consider the following problem:

1. B is on the right of A
2. C is on the left of B

3. D is in front of C
4. E is in front of B

What is the relation between D and E?

This problem supports the model:

```
C    A    B
D         E
```

But another model is compatible with this problem:

```
A    C    B
     D    E
```

Only one model is necessary to draw the conclusion that D is on the left of E because the alternative model yields the same conclusion. Nevertheless, this problem is labelled as a multiple-model problem. If the falsification procedure is seen as a necessity, then individuals who construct only one model are not "prudent" reasoners, reaching the correct conclusion accidentally (indeed, the alternative model could have refuted the putative conclusion supported by the first model, in particular if the premise had been "E is in front of A"). However, for the problem presented below the search for all possible models is not seen as a necessity for correct reasoning:

All the athletes are bakers
All the bakers are canoeists

This problem supports the model:

```
[[a]    b]    c
[[a]    b]    c
        ...
```

But other models are compatible with this problem, for example:

```
[[a]    b]    c
[[a]    b]    c
        [b]   c
              c
        ...
```

As in the case of the spatial problem, the alternative models do not refute the initial conclusion, that is, "all the athletes are canoeists." Nevertheless, this time the problem is considered a one-model problem (p. 107) and people constructing only the initial model are not viewed as imprudent. It seems, then, that sometimes the number of models associated with a problem – and consequently its difficulty – depends on the set of possible models, and sometimes it depends on the set of necessary models. So, do subjects attempt to construct the exhaustive set of models or do they attempt to construct only necessary models? One can try to answer by assuming that one will attempt to construct an alternative model if one presumes that this model contains information one considers relevant. That is, the model coveys information that could cause the revision of information reached initially and its construction does not exceed the maximum effort the individual is willing to expend (Sperber & Wilson 1995). Consider the description below:

A is taller than B
A is taller than C
A is taller than D
A is taller than E
A is taller than F.

With these premises, individuals will probably not try to construct any of the 120 possible models, first, because none of these models could express any determinate information other than that A is the tallest, and second, because the construction of all possible models will be very costly. One can then suggest that the falsification procedure is not only limited by working memory, as suggested initially by the authors (pp. 86 and 214), but also by expectations of relevance. This is what van der Henst's study (1999) tends to show (see also Schaecken et al. 1996).

Authors' Response

Mental models and pragmatics

P. N. Johnson-Laird[a] and Ruth M. J. Byrne[b]

[a]Department of Psychology, Princeton University, Princeton, NJ 08544;
[b]Department of Psychology, Trinity College, University of Dublin, Dublin 2, Ireland.
phil@clarity.princeton.edu www.tcd.ie/Psychology/People/Ruth_Byrne/
www.cogsci.princeton.edu/~phil/rmbyrne@tcd.ie

Abstract: Van der Henst argues that the theory of mental models lacks a pragmatic component. He fills the gap with the notion that reasoners draw the most relevant conclusions. We agree, but argue that theories need an element of "nondeterminism." It is often impossible to predict either what will be most relevant or which particular conclusion an individual will draw.

The theory of mental models postulates that individuals reason by envisaging the circumstances in which the premises and any other starting information are true (Johnson-Laird & Byrne 1991; 1993t). Each mental model represents a possibility, and so reasoners infer that a conclusion is possible if it holds in at least one model of the premises, that it is necessary if it holds in all the models of the premises, and that it is impossible if it holds in none of the models of the premises. Our book describing the theory and its corroboratory evidence received a mixed reception in *BBS.* Some commentators applauded the theory; some thought that it was – to a first approximation – rubbish. Nevertheless, the theory has flourished. Many authors have proposed interesting variants of the theory (e.g., Evans 1993; Polk & Newell 1995; Richardson & Ormerod 1997; Sloutsky & Morris, submitted). It has also been extended to reasoning based on suppositions (e.g., Byrne & Handley 1997), to temporal reasoning (e.g., Schaeken et al. 1996), to reasoning about probabilities (e.g., Johnson-Laird et al. 1999), and to counterfactual and causal reasoning (e.g., Byrne 1997; Goldvarg & Johnson-Laird, submitted).

A striking prediction of the theory is that certain inferences should be illusory. They are invalid yet most people should draw them, and they should seem compelling. They follow from the principle that mental models represent only those possibilities that are true given the premises, and a mental model represents the constituent propositions in the premises (affirmative or negative) only when they are true in the corresponding possibility. Consider, for example, the following problem (from Goldvarg & Johnson-Laird, in press):

Only one of the following assertions is true about a particular hand of cards:

> There is a king in the hand or there is an ace in the hand, or both.
> There is a queen in the hand or there is an ace in the hand, or both.
> There is a ten in the hand or there is a jack in the hand, or both.
> Is it possible that there is an ace in the hand?

Most people respond: yes (including 99% of the Princeton undergraduates whom we tested). Yet the response is an illusion. If there were an ace in the hand, then the first two assertions in the problem would be true, contrary to the rubric that only one assertion is true. Such illusions are a unique prediction of the model theory. They have been observed to occur in sentential reasoning (Johnson-Laird & Savary 1999), in quantified reasoning (Yang & Johnson-

Laird 1999), in probabilistic reasoning (Johnson-Laird et al. 1999), and in causal reasoning (Goldvarg & Johnson-Laird, submitted). Likewise, their predicted antidotes have also been corroborated (e.g., Newsome & Johnson-Laird 1996; Tabossi et al. 1999). If a theory is to be judged by the amount of work it has inspired, the model theory has survived its critique in *BBS.*

The theory is radically incomplete, but other theories of reasoning are no better. Indeed, a complete theory of reasoning would necessarily be a complete theory of the whole of cognitive psychology from perception to action. There are three principal gaps in the model theory. First, it gives no account of how general knowledge is represented in the mind or mobilized in reasoning. Second, the theory provides a semantics of standard connectives, temporal and spatial relations, and quantifiers, but it does not contain a full compositional semantics for a significant fragment of natural language, such as Basic English, or for various quantifiers that are not standard in formal logic, such as "most," "many," and "more than half." Third, the theory recognizes that the pragmatics of communication plays an important part in how people interpret the premises of inferences, and in what conclusions they are likely to draw. The theory, however, offers no account of pragmatics (Evans & Over 1997).

Van der Henst makes the same criticism, but he proposes to rectify matters by incorporating the theory of relevance, as formulated by Sperber and Wilson (1995). He points out that the theory of relevance makes powerful predictions about the particular conclusions that reasoners draw from premises. Because infinitely many valid conclusions follow logically from any set of premises, we had argued for several constraints. Reasoners tend to draw conclusions that maintain the semantic information in the premises, that express it parsimoniously, and that make explicit a proposition not explicitly asserted among the premises (Johnson-Laird & Byrne 1991). Van der Henst adds the further principle that reasoners tend to draw the most relevant conclusion in the context of processing. This principle is important, he says, when reasoners could draw several conclusions from the same premises. People are unlikely to reason just for fun, and so their goals – their reasons for reasoning – will be crucial in determining the conclusion that they draw.

Van der Henst illustrates his thesis with the following premises:

Paul is taller than John.
John is taller than Pete.
Pete is taller than Bob.

You can infer that Paul is the tallest, or Bob is the shortest, or that these individuals differ in height, depending on what is relevant to your goals. Likewise, even with indeterminate premises that support no valid conclusion interrelating the end terms, it may still be relevant to draw a conclusion. When you are interested in who won the race, and know only that Ann, Beth, and Cath were the leaders, you are more likely to draw a conclusion from premises of the form:

Ann came in ahead of Beth.
Ann came in ahead of Cath.

In the experiment that van der Henst describes, more participants drew the conclusion that Ann was the winner in this context than in a neutral context.

Van der Henst raises a second point concerning the search for alternative models. Given a spatially indeterminate description, we argued that people need to construct alternative models to reach the correct conclusion for the correct reasons. But, given premises of the form:

All the A are B.
All the B are C.

we postulated only a single mental model:

```
[[a]    b]    c
[[a]    b]    c,
```

which yields the valid conclusion: All the A are C. But the premises are consistent with other models, such as:

```
[[a]    b]    c
[[a]    b]    c
              c
```

Why do we not postulate that individuals construct this model? In fact, we have tinkered with the computer programs of the model theory of syllogistic reasoning, and in the latest version (described in Johnson-Laird & Byrne 1996), it does produce the preceding model. But, as we wrote in this article, our tinkering with the program now strikes us as a vain attempt to capture a highly flexible, if not labile, system within a single deterministic framework. Studies that externalize the search for alternative models suggest that people are indeed biased to construct a single model wherever possible, but the process of searching for counterexamples can be modeled only in a nondeterministic way (see Bucciarelli & Johnson-Laird 1999).

We agree with **van der Henst** on the importance of pragmatic issues. What is more controversial, however, is whether a reasoner's goals can be fully captured in Sperber and Wilson's (1995) conception of relevance. As we understand their theory, the relevance of an inference increases with its cognitive consequences and decreases with the amount of cognitive work needed to make the inference. These constraints are plausible, but sometimes difficult to put into practice, that is, to use to derive testable predictions. Moreover, as an unpublished study by van der Henst et al. shows, one person differs from another about what conclusion, if any, is relevant. In our view, theorists are unlikely to improve significantly on the sort of formulation offered by Sperber and Wilson, but they will be forced to allow for a considerable degree of "nondeterminism" in their theories. In many cases, it is impossible to predict which particular conclusion a particular individual will draw from particular premises.

References

Bucciarelli, M. & Johnson-Laird, P. N. (1999) Strategies in syllogistic reasoning. *Cognitive Science* 23:247–303. [rPNJ-L]

Byrne, R. M. J. (1997) Cognitive processes in counterfactual thinking about what might have been. In: *The psychology of learning and motivation. Advances in research and theory, vol. 37,* ed. D. L. Medin. Academic Press. [rPNJ-L]

Byrne, R. M. J. & Handley, S. J. (1997) Reasoning strategies for suppositional deductions. *Cognition* 62:1–49. [rPNJ-L]

Evans, J. St. B. T. (1993) The mental model theory of conditional reasoning: Critical appraisal and revision. *Cognition* 48:1–20. [rPNJ-L]

Evans, J. St. B. T. & Over, D. E. (1997) Rationality in reasoning: The problem of deductive competence. *Current Psychology of Cognition* 16:3–38. [rPNJ-L]

Goldvarg, Y. & Johnson-Laird, P. N. (submitted) Naive causality: A mental model theory of causal meaning and reasoning. (1999a). [rPNJ-L]

(in press) Illusions in modal reasoning. *Memory and Cognition.* (1999b). [rPNJ-L]

Johnson-Laird, P. N. & Byrne, R. J. M. (1991) *Deduction.* Erlbaum. [J-BvdH, rPNJ-L]

(1993) Précis of *Deduction. Behavioral and Brain Sciences* 16:323–80. [J-BvdH, rPNJ-L]
(1996) Authors' reply to Hardman: Mental models and syllogisms. *Behavioral and Brain Sciences* 19:543–46. [rPNJ-L]
Johnson-Laird, P. N., Legrenzi, P., Girotto, V., Legrenzi, M. & Caverni, J.-P. (1999) Naive probability: A mental model theory of extensional reasoning. *Psychological Review* 106:62–88. [rPNJ-L]
Johnson-Laird, P. N. & Savary, F. (1999) Illusory inferences: A novel class of erroneous deductions. *Cognition* 71:191–229. [rPNJ-L]
Newsome, M. R. & Johnson-Laird, P. N. (1996) An antidote to illusory inferences. *Proceedings of the Eighteenth Annual Conference of the Cognitive Science Society,* p. 820. Erlbaum. [rPNJ-L]
Polk, T. A. & Newell, A. (1995) Deduction as verbal reasoning. *Psychological Review* 102:533–66. [rPNJ-L]
Richardson, J. & Ormerod, T. C. (1997) Rephrasing between disjunctives and conditionals: Mental models and the effects of thematic content. *Quarterly Journal of Experimental Psychology* 50A:358–85. [rPNJ-L]
Schaecken, W., Johnson-Laird, P. N. & d'Ydewalle, G. (1996) Mental models and temporal reasoning. *Cognition* 60:205–34. [J-BvdH, rPNJ-L]
Sloutsky, V. M. & Morris, B. J. (submitted) How to make something out of nothing: Adaptive constraints on children's information processing. [rPNJ-L]
Sperber, D. & Wilson, D. (1987) Précis of *Relevance. Behavioral and Brain Sciences* 10:697–754. [J-BvdH]
(1995) *Relevance: Communication and cognition, 2nd edition.* Blackwell. [J-BvdH, rPNJ-L]
Tabossi, P., Bell, V. A. & Johnson-Laird, P. N. (1999) Mental models in deductive, modal, and probabilistic reasoning. In: *Mental models in discourse processing and reasoning,* ed. C. Habel & G. Rickheit. John Benjamins. [rPNJ-L]
van der Henst, J.-B. (1999) The mental model theory and spatial reasoning re-examined: The role of relevance in premise order. *British Journal of Psychology.* (in press). [J-BvdH]
van der Henst, J.-B., Politzer, G. & Sperber, D. (2000) In search of a relevant conclusion: A pragmatic analysis of indeterminate relational problems. (in preparation). [J-BvdH]
Yang, Y. & Johnson-Laird, P. N. (1999) Illusory inferences with quantified assertions. *Memory and Cognition.* (in press). [rPNJ-L]

Commentary on Tracey J. Shors & Louis D. Matzel (1997) Long-term potentiation: What's learning got to do with it? BBS 20:597–655.

Abstract of the original article: Long-term potentiation (LTP) is operationally defined as a long-lasting increase in synaptic efficacy following high-frequency stimulation of afferent fibers. Since the first full description of the phenomenon in 1973, exploration of the mechanisms underlying LTP induction has been one of the most active areas of research in neuroscience. Of principal interest to those who study LTP, particularly in the mammalian hippocampus, is its presumed role in the establishment of stable memories, a role consistent with "Hebbian" descriptions of memory formation. Other characteristics of LTP, including its rapid induction, persistence, and correlation with natural brain rhythms, provide circumstantial support for this connection to memory storage. Nonetheless, there is little empirical evidence that directly links LTP to the storage of memories. In this target article we review a range of cellular and behavioral characteristics of LTP and evaluate whether they are consistent with the purported role of hippocampal LTP in memory formation. We suggest that much of the present focus on LTP reflects a preconception that LTP *is* a learning mechanism, although the empirical evidence often suggests that LTP is unsuitable for such a role. As an alternative to serving as a memory storage device, we propose that LTP may serve as a neural equivalent to an arousal or attention device in the brain. Accordingly, LTP may increase in a nonspecific way the effective salience of discrete external stimuli and may thereby facilitate the induction of memories at distant synapses. Other hypotheses regarding the functional utility of this intensely studied mechanism are conceivable; the intent of this target article is not to promote a single hypothesis but rather to stimulate discussion about the neural mechanisms underlying memory storage and to appraise whether LTP can be considered a viable candidate for such a mechanism.

LTP – A mechanism in search of a function

Kathryn J. Jeffery
Department of Anatomy and Developmental Biology, University College London, London WC1E 6BT, United Kingdom. **kate@maze.ucl.ac.uk**

Abstract: Shors & Matzel (1997) suggest replacing the question "Is LTP a mechanism of learning?" with "Is LTP a mechanism of arousal and attention?" However, the failure of experiments to verify the LTP-learning hypothesis may arise not because it is untrue, but because in its current guise, it is not properly testable. If so, then the LTP-attention hypothesis is untestable, as well.

The hypothesis that links LTP to the mechanisms of learning is now a quarter of a century old. Shors & Matzel's (S&M's) (1997) target article, a broad-reaching and well-written review of the evidence to date, argues persuasively that there is still no reason either to accept or reject it outright. The response from the commentators supports this uncertainty, some agreeing that support for the hypothesis is indeed weak, others arguing that the right experiments have not yet been done. This leaves behavioural physiologists in something of a quandary. Should we press on, continuing to try many and various different ways of tying the two phenomena together for perhaps another quarter of a century, or should we follow S&M's advice and abandon the learning hypothesis, replacing it instead with something new, such as arousal or attention?

Clearly, the current approach to tackling the LTP-learning question has been unsuccessful in resolving the question and so the answer to the first question is "no." However, we should look carefully at the underlying reasons before we make the mistake of stumbling down another, equally stony path of investigation in pursuit of the neurobiological mechanisms of arousal. That so much hard work and so many experiments have failed to confirm a hypothesis that remains widely believed should raise a warning flag that it might be not the hypothesis but rather the means of testing it that is flawed.

At this point it is worth reiterating the well-worn point that LTP is an *experimental* phenomenon. Its study has uncovered some intricate synaptic machinery that probably does exist to change connection strengths between neurons. However, we should not make the mistake of confusing the question of what this synaptic modifiability does for an animal with the (methodologically easier) question of what *LTP* does for an animal. To keep this point in the foreground, therefore, it is useful to distinguish between LTP, on the one hand, and the putative phenomenon of naturally occurring synaptic modification (SM) on the other. LTP has been put forward as a model of naturally occurring SM, but it is *not the same thing.* Therefore, the question "Is LTP a mechanism of learning?" is really two questions: (1) Does SM underlie learning? and (2) Is LTP a good model of SM?

The study of model systems like LTP can be a useful tool in neurobiology, because it enables experimenters to isolate the phenomenon of interest and explore it in the laboratory. However, when study of a model fails to confirm a hypothesis, it may be that the hypothesis is wrong, but it may also simply be that the model is unsuitable. In the case of LTP, Shors and Matzel argue that the hypothesis is wrong and we should therefore find a different one. However, it is also possible that the hypothesis (that SM is the mechanism of learning) is correct but the model (LTP in a given pathway) is wrong. For example, it may be that the synaptic changes of LTP are not identical to those of SM, the differences contributing to the experimental results. Perhaps the pathway in which LTP was evoked was not that involved in the learning of the task, or perhaps the method of inducing LTP (with theta-burst or paired or tetanic or primed burst stimulation or whatever) did not mimic naturally occurring conditions. The list goes on.

The inevitable conclusion is that although the SM-learning hypothesis might have been supported by a large number of positive correlations between the properties of LTP and those of learning, a failure to find such correlations, or at least to find them reliably, cannot be construed as evidence that the hypothesis is wrong. This is because not enough is known about the system we are investigating to know whether LTP is a good model of it. It follows from this argument that the worst possible course of action would be to throw into the pool yet another hypothesis, about a process that is even less understood and less well localised than learning. There is no point in using LTP as a model for arousal, or anything else, if the process it is supposed to model has not even been partially characterised.

How better to characterise learning? The top-down approach would be to break it up into its simplest components and find out where in the brain these occur, using pharmacological and lesion techniques. The bottom-up approach would be to observe the behaviour of single neurons to see what they actually do when learning occurs. This means knowing what the neurons represent, what the animal learned, what the neurons learned (and for an individual neuron, this may not be the same thing), and whether the cell-to-cell communications changed after the process occurred. If the learning event involved a change in the connection strength between a pair of neurons, then, and *only* then, should our wealth of knowledge about LTP be brought into play.

In short, then, we should not throw away LTP as a model of learning-related synaptic change until its suitability has been discredited. Rather, we should set it aside while we better characterise the processes underlying learning, and this means discovering (a) where they happen, and (b) under what conditions. Until synaptic strength changes can be observed to participate directly in a given process, any attempts to postulate an underlying LTP-like mechanism can only be speculative.

LTP and reinforcement: Possible role of the monoaminergic systems

Mikhail N. Zhadin

Laboratory of Neurocybernetics, Institute of Cell Biophysics, 142292 Pushchino, Russia. **zhadin@online.stack.net**

Abstract: The absence of a clear influence of the responses modified by new connections created by LTP on the development of these connections casts doubt on an essential role of LTP in learning and memory formation without any association with reinforcement. The evidence for the involvement of the monoaminergic systems in synaptic potentiation in the cerebral cortex during learning is adduced, and their role in reinforcement system function is discussed.

I share Shors & Matzel's (1997t) doubts about the role of long-term potentiation (LTP) in memory trace formation and with Latash's (1997) more radical criticism of the concept of LTP as a key function in memory trace formation. It seems rather strange that this idea arose at all and persisted so long without any association with the concept of reinforcement. Some further considerations support the above doubts.

As a result of evolution, memory has developed as a mechanism of adaptation to the environment in which an animal must choose the behavioral responses useful for its survival and avoid making harmful ones. The memory trace, consisting of new connections in the cortex, is formed under the influence of continually arriving information about the consequences of the animal's behavior, together with an estimate of its usefulness or harmfulness. So there must be mutual relations between new connection formation in the cerebral cortex and the animal's behavior: The potentiation of available connections and the formation of new ones influence the behavioral responses, and the estimate of the results of the animal's behavior in turn exerts an effect on the formation of new connections and the modulation of existing ones. In the event of LTP, however, relations of this sort are only unidirectional. LTP creates new connections, acting on the animal's behavior, but any obvious mechanism for the reverse influence of these new or modified responses on forming the above connections is absent. Therefore, because these connections occur independently of the maintenance of the animal's optimal behavior, it seem highly improbable that LTP is important in memory trace formation.

According to the general synaptic theory of learning and memory, as well as the physiology of higher nervous activity, it is reinforcement that provides brain structures (including the cerebral cortex) with information about the usefulness or harmfulness of each behavioral response. The great diversity of learning networks (e.g., McCulloch & Pitts 1943; Rosenblatt 1962) constructed on this basis have been able to elaborate the requisite reactions in response to rather complicated forms of conditioned signals.

This raises the question: What transmits the information about the consequence of an animal's behavior to its cortex and influences synaptic potentiation in the cerebral cortex at learning? A possible participation of the monoaminergic systems of the brain was postulated in the middle 1970s (Freedman et al. 1977; Libet et al. 1975; Zhadin 1977).

Support comes from a variety of evidence:

(1) the association of monoaminergic nucleus activity with the reinforcement system (Gromova 1980);

(2) the diffuse distribution of monoaminergic fibers from certain local nuclei throughout the brain, in particular, over the whole cortex (Ungerstedt 1971);

(3) the unique chemistry of monoaminergic synapses, not found among cortico-cortical synapses or among internal connections in other structures of the brain (Ungerstedt 1971);

(4) the difference between the monoaminergic systems hypothetically associated with positive and negative reinforcement (Gromova 1980);

(5) the special structure of the monoaminergic synapses in the cerebral cortex; they have only presynaptic components, without any postsynaptic ones, and saturate all the cortical extracellular space with corresponding monoamine upon excitation of these synapses, providing an easy access to the monoamines for all the cortico-cortical synapses (Beaudet & Descarries 1978).

This all provides grounds to believe that it is the monoaminergic systems that promote the proper changes in efficacy of cortico-cortical synapses depending on the previous state of these synapses for the time period between the conditioned stimulus and response, as well as on the quality of reinforcement (positive or negative) at learning (Zhadin 1987; 1991; 1993).

According to our mathematical model (Zhadin 1987; 1991; 1995; Zhadin & Bakharev 1987), the prolonged action of positive reinforcement induces gradual transitions in relatively highly excited neurons in the cortex to their extremely high excitation levels and of less excited ones to complete inhibition, with occasional possible spontaneous transitions of the cells from one of these extreme levels to the other. The long-term negative reinforcement causes gradual transitions of both the relatively active and less excited neurons to some intermediate level of neuronal excitation and subsequent stabilization at that level.

Our experiments (Mamedov 1987; Zhadin 1987) with long-term application of monoamines to the intact cortex have shown that serotonin induces effects characteristic of the positive reinforcement, and noradrenaline those of negative reinforcement. Serotonin, applied to the completely deafferented neocortex together with acetylcholine activating the cortex, also caused a rise in the transition from inhibition to extremely high neuronal excitation, and vice versa (Ignat'ev et al. 1986). This provided evidence for the appropriate changes in synaptic connections between cortical neurons. The experimental and theoretical results suggested that the serotoninergic system mediates positive reinforcement and the noradrenergic system mediates negative reinforcement. The experiments (Zhadin & Karpuk 1996) on neocortical surviving slices showed that introducing serotonin into the incubating medium strengthened interaction between relatively highly excited neurons and reduced it between initially less active cells.

Under the influence of these monoamines changes in synaptic efficacy took place at the neuronal frequencies typical for the neocortex: in the range up to 10 impulses/sec; the frequencies of the order of 100/sec required for the LTP development are not normally encountered among neocortical neurons. The situation with the synaptic potentiation in learning accordingly seems to be more complicated than simple LTP.

Authors' Response

The status of LTP as a mechanism of memory formation in the mammalian brain

Tracey J. Shors[a] and Louis D. Matzel[b]

[a]Department of Psychology and Center for Neuroscience, Rutgers University, New Brunswick, NJ 07728; [b]Department of Psychology, Program in Biopsychology and Behavioral Neuroscience, Rutgers University, New Brunswick, NJ 08903. **shors@rci.rutgers.edu** **matzel@rci.rutgers.edu**

Abstract: Long-term potentiation (LTP) is a long-lasting increase in synaptic efficacy that many consider the best candidate currently available for a neural mechanism of memory formation and/or storage in the mammalian brain. In our target article, *LTP: What's learning got to do with it?*, we concluded that there was insufficient data to warrant such a conclusion. In their commentaries, **Jeffery** and **Zhadin** raise a number of important issues that we did not raise, both for and against the hypothesis. Although we agree with a number of these issues, we maintain that there remains insufficient evidence that LTP is a memory mechanism.

In our target article (Shors & Matzel 1997; 1997r), we catalogued the plurality of empirical evidence that leads to the logical imperative that long-term potentiation (LTP), and in particular, N-methyl-D-aspartate (NMDA) receptor-dependent forms of hippocampal LTP, is poorly suited to serve as a substrate mechanism for memory storage. Although in isolation, certain results and some of the neurophysiological characteristics of LTP provide a degree of support for the common assertion that LTP *is* a memory-storage device. Converging evidence, as well as fundamental incompatibilities, between mechanistic features of LTP and the process of memory led us to conclude that LTP is poorly suited for this function. In what we considered a simple illustration (and but one example) of the equivocal nature of the data to date, we argued that the extant evidence linking LTP to memory could be similarly interpreted as indicative of a role for LTP in the modulation of attention (and indirectly, memory). This line of reasoning impressed on us (as we hoped it would the reader) the inadequacy of prevailing approaches to this question. If a single set of data could be interpreted as support for two very different hypotheses, then the data could not be considered conclusive. To our surprise (and regret), many of the commentaries on our target article focused on the inadequacy of our alternative hypothesis. In some instances, the rejection of our alternative hypothesis was deemed a confirmation of the assertion that LTP is a memory-storage device. It is unfortunate that the presentation of an alternative hypothesis regarding the physiological function of LTP distracted so many readers from the larger issue raised, namely, that the "fit" between data purported to support a link between LTP and memory storage was far from compelling. Definitive statements to the contrary were the reflection of popular consensus, not empirical evidence.

Consistent with a general reluctance of some to consider alternative roles for LTP, **Jeffery** argues that it is currently beyond our capability to consider neurophysiological mechanisms for attention, because as she asserts, attention is more poorly characterized than learning. Of course, our lack of understanding is not relevant to the function of a biological system. Despite her acknowledgment that there is no compelling evidence to support the conclusion that LTP serves as a memory storage device, she argues that the investment that has already been made should compel us to continue along similar lines. Again, our investment is not relevant to the function of a biological system.

Instead of focusing on new roles for LTP (as we did), **Jeffery** suggests (as we did) that other approaches are necessary to address the issue definitively. In this regard, she notes that there is no compelling evidence that synaptic modifications play any direct role in the storage of memories. Prior to further exploration of the role of LTP in memory (which is but one form of a plethora of mechanisms by which synapses are modified), she suggests that investigators first attempt to confirm that memory storage is the product of *any* form of change in synaptic efficacy. Although it is almost universally accepted that specific modifications of synaptic efficacy regulate the induction of memory, direct *empirical* support for this immensely popular assumption is nonexistent. (As an aside, contrast Jeffery's observation with that of P. M. Milner who referred to the "overwhelming" evidence that memory storage was the product of "changing the effectiveness of synapses.") If there were even one synapse or set of synapses in the vertebrate nervous system *known* to be modified by experience *and essential* to the expression of a memory for that experience, it would only be a practical matter to determine whether that specific synaptic modification followed rules established for the induction of LTP. There is a reason that this simple and definitive experiment has not been done. Preconceptions and popular conviction aside, a single convincing (or even strongly suggestive) example of such a synaptic modification is yet to be found in the vertebrate brain. Although it was not the point we had intended to make in our target article, we agree with Jeffery that there is little direct evidence, if any, that synapse-specific modifications are the substrate for memory in the vertebrate brain (cf. Matzel et al. 1998).

Jeffery suggests that we should first determine whether synaptic modification underlies learning, and only then begin to evaluate the appropriateness of LTP as a model for synaptic modification. Jeffery's reasoning is analogous to our concern that in many cases, LTP has become nothing more than a synonym for synaptic facilitation. As such, it has become such a nebulous concept that it can never be disconfirmed, regardless of the aggregate of seemingly antithetical data. It is in this regard that Jeffery argues that evidence to date neither confirms nor disconfirms the hypothesis that LTP is a memory storage device. Although we understand and even sympathize with the argument, we would contend that there are *prerequisite* requirements that a mechanism must meet before it should be considered for this role. For example, if memories can persist intact for the lifetime of an organism, then the memory storage device should be comparably enduring. Although Jeffery notes that confirmatory evidence for the role of LTP in memory is not yet available, we do know that LTP decays *rapidly* (in hours or days). This fact alone suggests that it is a poor candidate for a memory storage device. More or better experiments will not change this immutable characteristic of the mechanism. Furthermore, there are numerous instances in which learning is present despite explicit impairment of LTP. In mice lacking the cell adhesion molecule Thy-1, it was observed that LTP could not be induced under optimal conditions *in vitro,* yet learning was intact (Nosten-Bertrand et al. 1996). It was later determined that a depressed form of LTP could be induced in these mice when assessed in an *in vivo* preparation (Errington et al. 1997), leading the authors to suggest that both *in vitro* and *in vivo* recordings were necessary to assess accurately LTP's relation to learning. Of course, these caveats do not negate the finding that LTP and learning were dissociated.

Related to the issue of dissociations between LTP and memory, it was recently reported that the initial saturation study of Castro et al. (1989), in which prior induction of LTP occluded subsequent learning, had finally been replicated. In the initial study, it was reported that saturation of LTP in the dentate gyrus, a brain region known to express LTP and to be involved in some types of memory, impaired spatial memory. But others were later unable to replicate this finding (Cain et al. 1993; Korol et al. 1993; Robinson 1992; Sutherland et al. 1993). Moser et al. (1998) were able to replicate at least partially the initial work of Castro et al. However, to observe any effect of saturation, unilateral lesions of the nontetanized hippocampus were necessary, as was massive bundle tetanization of the remaining hippocampus. Thus the functional lesion induced by saturation of one hippocampus was without effect if not combined with a neurotoxic lesion of the remaining hippocampus! Note that *simply* tetanizing both hippocampi was without effect. It is unfortunate that such heroic efforts are not directed toward experiments intended to disprove the hypothesis that LTP is not the memory mechanism.

Although we are encouraged by **Jeffery**'s skeptical view of the data purported to support the link between LTP and memory storage, we are less enthusiastic about the second general point she makes regarding the impediments imposed by the complexity of attention or the effort already invested in LTP as a memory mechanism. It is entirely unproductive to suggest that we maintain an exclusive emphasis on the elucidation of LTP's role in memory because its potential role in other processes might be too difficult to establish or because of the time and energy already committed to the dominant hypothesis. As mentioned, prior investment is no rationale for adhering to an untenable hypothesis, and the difficulty associated with testing a hypothesis is unrelated to the validity of that hypothesis. For reasons outlined in our target article, we believe that the evidence for LTP's function as a gain control (or attentional) device is more compelling than the more popular alternative hypothesis regarding LTP's role in memory. Given that the available data does not adequately distinguish between the two hypotheses is all the *more* reason to devise experiments that do. Considering the more than 30 memory systems alluded to in our target article, we are not convinced that attention is less well understood than memory as Jeffery suggests, but we do agree that attention is a difficult concept to operationalize. Nevertheless, the complexity of a process should not dissuade us from considering it. Experimental programs designed to distinguish among alternative hypotheses are the means by which we converge on the correct alternative.

Zhadin suggests that one of the reasons for our lack of progress in understanding LTP's role in memory is the less-than-rigorous interpretation of fundamental learning and memory processes. For example, associative LTP is optimally induced when pre- and postsynaptic activity occurs with a forward interstimulus interval (ISI) of less than 100 msec. This temporal constraint on LTP induction is commonly cited as cardinal evidence for the role of LTP in the storage of associative memories, because a "general" feature of associative learning is that it is optimally induced when the conditioned stimulus (CS) and unconditioned stimulus (US) are presented with a forward relationship and a "short" ISI. Though consistent with the most rudimentary views of associative learning, this assertion is quite simply wrong as it pertains to this instance. In contrast to the narrow temporal window with which associative LTP is induced, there is no "optimal" ISI for the induction of associative memories. Rather, the efficacy of an ISI is determined by the conditions under which the animal is trained. The effective range of ISIs varies across conditioning paradigms, as a function of the response being measured, and the nature and modality of the stimuli to be associated. Even so, the ISI that supports eye blink conditioning (the shortest of any common conditioning preparation) is still hundreds of msecs *longer* than the ISI required for successful induction of associative LTP. In contrast, fear conditioning and heart rate conditioning are optimally induced with ISIs longer than 5 sec, and at the furthest extreme, taste aversion learning is actually impaired with ISIs of less than several hours and can be easily attained with ISIs of 24 hours! Thus the premise that "the" ISI in associative learning could be determined by the temporal constraints on associative LTP is based on a view of associative learning that is wholly incompatible with the most fundamental properties of this process.

Whereas the temporal constraint on associative LTP was once eagerly and broadly embraced as evidence for its role in associative learning, this early enthusiasm has given way to a more critical appraisal of its limitations. However, rather than abandon such an unworkable premise, these reappraisals have fostered yet another series of efforts to fit the square peg of LTP into the round hole of learning. For example, some investigators have proposed that a cell's latency to fire an action potential imposes the actual interval between pre- and postsynaptic activity, or the ISI (Faulkner & Brown 1999). One need only assume that the latency to fire in *some* cells that process the CS correspond to the optimal ISI, which supports learning under any given set of training conditions. If this were the case, it is proposed that LTP could still serve as the "coincidence detector" underlying the formation of CS-US associations. However, since a single stimulus (CS) can serve effectively to predict any one of a virtually infinite number of USs, and since the US (and the unconditioned response to it) determines the ISI that will support optimal learning (or conditioned responding). Thus, frameworks such as these, which still assign to LTP the task of a "coincidence detector," are as untenable as their more simplistic predecessors. For their effective application, these more elaborate variations on the old premise require that the ISI be a preestablished property of some population of cells that process the CS, and as such, would require a unique population of such cells for any of our infinite number of CS-US diads.

Although such a property of a network of cells could support learning under a limited set of circumstances, to be more generally applicable such a system requires that a nervous system maintain a specific population of cells for every *potential* CS-US dyad that the animal might ever encounter. Such an allocation of resources, if physically possible, would be enormously inefficient. Although these explanations of association formation are mechanistically limited in their application to only a specific set of conditions, the capacity for learning by real animals is not so severely constrained. Nature did not design a world in which only specific stimuli could co-occur, and so evolution provided animals with nervous systems capable of making associations between functionally infinite combinations of stimuli.

In addition to the adaptive and flexible nature of the ISI during associative conditioning, association formation is not limited to instances where the CS precedes the US. Associations can also be formed between a CS and US in instances where the CS is presented in a simultaneous or backward relationship to the US. Modern conceptualizations of associative learning provide that although the directional relationship between stimuli arbitrarily designated as the CS and US may determine the form of the conditioned response, it does not necessarily limit the animal's ability to learn. Yet, if associative LTP were the mechanism underlying association formation, only forward conditioning would be possible. In this regard, **Zhadin** is entirely right in his assertion that the inherent nature of LTP limits its possible application to the coding of unidirectional relationships, that is, it would provide an animal with the capacity to learn but a single relationship between two events and prohibit the potential to respond adaptively to a stimulus with multiple meanings or in response to a change in the prevailing conditions.

Although we share **Zhadin**'s more general concern about the lack of evidence associating LTP and memory, we disagree with his specific assertion that LTP fails as a potential learning device because it does not provide a means for feedback from the reinforcer that supports learning. We see no reason why a single synapse or set of synapses must be bi-directionally regulated by reinforcement when different sets of synapses can independently code an association and the response that is appropriate for it. Moreover, Zhadin makes an error similar to the one that he objects to, that is, he supposes that an effective learning device should modify a reflex arc (in this case bi-directionally) so that an animal can alter its response repertoire to accommodate a particular reinforcer. We see no need to so constrain the process of learning. As Rescorla (1988) stresses, the function of associative learning is to provide an animal with knowledge about stimulus *relationships,* not simply to modify an evoked *response.* Furthermore, common sense might suggest that learning about relationships between nominally neutral stimuli (i.e., without "reinforcement" value) can be as integral to an animal's capacity to negotiate its environment as learning about nominal reinforcers. The capacity of animals to associate neutral stimuli has long been recognized (e.g., Brogden 1939; Matzel et al. 1988; Rescorla 1980). Although we agree with Zhadin that the mechanism of LTP could describe the intractable modification of reflexes, we disagree with his contention that learning can only be described by a mechanism that can process feedback from "reinforcers." In the end, we must disagree with Zhadin's notion that there must be "mutual relations between new connection formation in the cerebral cortex and the animal's behavior." Memory is not necessarily synonymous with behavior. Behavior is simply one of our dependent measures for assessing learning, but is not a prerequisite or a necessary consequence.

In our target article (Shors & Matzel 1997t), we encouraged the generation of new and testable hypotheses, not only for LTP's function in memory, but also for mechanisms of memory formation itself. **Zhadin** has postulated that monoaminergic systems would be one means for achieving enhanced synaptic efficacy. He goes on to propose that serotonergic systems mediate "positive" and noradrenergic systems mediate "negative" reinforcement. It strikes us as odd to suppose that a transmitter system would evolve for the specific purpose of coding stimuli based on our qualitative characterization of them. As discussed, reinforcement is not a necessary prerequisite for learning. In addition, however, the proposal for monoaminergic modulation of synaptic efficacy does not add evidence one way or the other to the premise that LTP is a memory storage device. It simply assumes that LTP is the memory storage mechanism that can be modulated by neurotransmitters other than glutamate. In our target article we eluded to the nearly infinite list of modulators of LTP, including monoamines. Further characterization of the modulation of LTP will not reveal whether LTP is a memory mechanism.

Although we disagree with several of the specific issues raised by both **Jeffery** and **Zhadin,** we share their more general concern that the evidence supporting a role for LTP in memory storage is far from compelling. We are encouraged in this regard, and are pleased to see others question this early consensus. Minimally, these critical but divergent appraisals of the LTP-memory hypothesis are further confirmation of our principle contention that, despite the prevalence of assertions to the contrary, objective observers can come to no consensus regarding what role, if any, LTP plays in the induction or storage of memories.

References

Beaudet, A. & Descarries, L. (1978) The monoamine innervation of rat cerebral cortex: Synaptic and nonsynaptic axon terminals. *Neuroscience* 3:851–60. [MNZ]

Brogden, W. J. (1939) Sensory pre-conditioning. *Journal of Experimental Psychology* 25:323–32. [rTJS]

Cain, D. P., Hargreaves, E. L., Boon, F. & Dennison, Z. (1993) An examination of relations between hippocampal long-term potentiation, kindling, afterdischarge, and place learning in the water-maze. *Hippocampus* 3:153–64. [rTJS]

Castro, C., Silbert, L., McNaughton, B. & Barnes, C. (1989) Recovery of spatial learning deficits after decay of electrically induced synaptic enhancement in the hippocampus. *Nature* 342:545–48. [rTJS]

Errington, M. L., Bliss, T. V. P., Morris, R., Laroche, S. & Davis, S. (1997) Long-term potentiation in awake mutant mice. *Nature* 387:666–67. [rTJS]

Faulkner, B. & Brown, T. H. (1999) Morphology and physiology of neurons in the rat perirhinal-lateral amygdala area. *Journal of Comparative Neurology* 411:613–42. [rTJS]

Freedman R., Hoffer, R. J., Woodward, D. J. & Puro, D. (1977) Interaction of norepinephrine with cerebellar activity evoked by mossy and climbing fibers. *Experimental Neurology* 55:269–88. [MNZ]

Gronova, E. A. (1980) Emotional memory and its mechanisms. *Science* (in Russian). [MNZ]

Ignat'ev, D. A., Agladze, N. N. & Zhadin, M. N. (1986) Effect of serotonin and acetylcholine on electrical activity of the isolated rabbit cortex. *Neuroscience and Behavioral Physiology* 16:376–83. [MNZ]

Korol, D., Abel, T., Church, L., Barnes, C. & McNaughton, B. (1993) Hippocampal synaptic enhancement and spatial learning in the Morris swim test. *Hippocampus* 3:127–32. [rTJS]

Latash, L. P. (1997) LTP is neither a memory trace nor an ultimate mechanism for its formation: The beginning of the end of the synaptic theory of neural memory. *Behavioral and Brain Sciences* 20:621–22. [MNZ]

Libet, B., Kobayashi, H. & Tanaka, T. (1975) Synaptic coupling into the production and storage of a neuronal memory trace. *Nature* 258:155–57. [MNZ]

Mamedov, Z. G. (1987) Changes in the activity of cortical neurons under the influence of biogenic amines. *Neuroscience and Behavioral Physiology* 17:160–67. [MNZ]

Matzel, L. D., Held, F. P. & Miller, R. R. (1988) Information and the expression of simultaneous and backward associations. *Learning and Motivation* 19:317–44. [rTJS]

Matzel, L. D., Talk, A. C., Muzzio, I. & Rogers, R. F. (1998) Ubiquitous molecular substrates for associative learning and activity-dependent neuronal facilitation. *Reviews in Neuroscience* 9:129–67. [rTJS]

McCulloch, W. S. & Pitts, W. (1943) A logical calculus of the ideas immanent in the nervous system. *Bulletin of Mathematical Biophysics* 5:115–33. [MNZ]

Moser, E., Krobert, K. A., Moser, M. & Morris, R. G. (1998) Impaired spatial learning after saturation of long-term potentiation. *Science* 281:2038–42. [rTJS]

Nosten-Bertrand, M., Errington, M. L., Murphy, K. P., Tokugawa, Y., Barboni, E., Koslova, E., Michalovich, D., Morris, R. G., Silver, J., Stewart, C. L., Bliss, T. V. P. & Morris, R. J. (1996) Normal spatial learning despite regional inhibition of LTP in mice lacking Thy-1. *Nature* 379:826–29. [rTJS]

Rescorla, R. A. (1980) Simultaneous and successive associations in sensory preconditioning. *Journal of Experimental Psychology: Animal Behavior Processes* 6:207–16. [rTJS]

(1988) Pavlovian conditioning. It's not what you think it is. *American Psychologist* 43:151–60. [rTJS]

Robinson, G. (1992) Maintained saturation of hippocampal long-term potentiation does not disrupt acquisition of the eight-arm radial maze. *Hippocampus* 2:389–95. [rTJS]

Rosenblatt, F. (1962) *Principles of neurodynamics. Perceptions and the theory of brain mechanisms.* Spartan. [MNZ]

Shors, T. J. & Matzel. L. D. (1997t) Long-term potentiation (LTP): What's learning got to do with it? *Behavioral and Brain Sciences* 20:597–613. [KJJ, rTJS, MNZ]

(1997r) LTP: Memory, arousal, neither, both. *Behavioral and Brain Sciences* 20:634–44. [rTJS]

Sutherland, R. J., Dringenberg, H. C. & Hoesing, J. M. (1993) Induction of long-term potentiation at perforant-path dentate synapses does not affect place learning or memory. *Hippocampus* 3:141–48. [rTJS]

Ungerstedt, U. (1971) Stereotaxic mapping of the monoamine pathway in the rat brain. *Acta Physiologia Scandinavica* (Suppl.) 367:1–48.

Zhadin, M. N. (1977) Model of conditioned reflex formation and analysis of functional significance of electrophysiological correlates of learning. *Journal of Higher Nervous Activity* 27:949–56. (in Russian). [MNZ]

(1987) Electrophysiological manifestations of the monoaminergic systems' effects on the cerebral cortex. *Neuroscience and Behavioral Physiology* 17:152–60. [MNZ]

(1991) Biophysical mechanisms of the EEG formation. In: *Mathematical approaches to brain functioning diagnostics,* ed. I. Dvorak & A. Holden. Manchester University Press. [MNZ]

(1993) Possible mechanism of the action of biogenic amines on the activity of cortical neurones. *Biophysics* 38:353–58. [MNZ]

(1995) Collective behaviour of cortical neurons on prolonged exposure to reinforcement. *Biophysics* 40:631–34. [MNZ]

Zhadin, M. N. & Bakharev, B. V. (1987) Model of variations in the level of cortical neuron excitation at increased biogenic amine concentration. *Studia Biophysica* 121:81–88. [MNZ]

Zhadin, M. N. & Karpuk, N. N. (1996) Influence of serotonin on cross-correlation in neuronal activity in surviving slices of the cerebral cortex. *Journal of the Higher Nervous Activity* 46:547–51. (in Russian). [MNZ]

Commentary on **Siu L. Chow (1998). Précis of *Statistical significance: Rationale, validity, and utility.* BBS 21:169–239.**

Abstract of the original article: The null-hypothesis significance-test procedure (NHSTP) is defended in the context of the theory-corroboration experiment, as well as the following contrasts: (a) substantive hypotheses versus statistical hypotheses, (b) theory corroboration versus statistical hypothesis testing, (c) theoretical inference versus statistical decision, (d) experiments versus nonexperimental studies, and (e) theory corroboration versus treatment assessment. The null hypothesis can be true because it is the hypothesis that errors are randomly distributed in data. Moreover, the null hypothesis is never used as a categorical proposition. Statistical significance means only that chance influences can be excluded as an explanation of data; it does not identify the nonchance factor responsible. The experimental conclusion is drawn with the inductive principle underlying the experimental design. A chain of deductive arguments gives rise to the theoretical conclusion via the experimental conclusion. The anomalous relationship between statistical significance and the effect size often used to criticize NHSTP is more apparent than real. The absolute size of the effect is not an index of evidential support for the substantive hypothesis. Nor is the effect size, by itself, informative as to the practical importance of the research result. Being a conditional probability, statistical power cannot be the *a priori* probability of statistical significance. The validity of statistical power is debatable because statistical significance is determined with a single sampling distribution of the test statistic based on H_0, whereas it takes two distributions to represent statistical power or effect size. Sample size should not be determined in the mechanical manner envisaged in power analysis. It is inappropriate to criticize NHSTP for nonstatistical reasons. At the same time, neither effect size, nor confidence interval estimate, nor posterior probability can be used to exclude chance as an explanation of data. Neither can any of them fulfill the nonstatistical functions expected of them by critics.

Statistical significance testing, hypothetico-deductive method, and theory evaluation

Brian D. Haig

Department of Psychology, University of Canterbury, Christchurch, New Zealand. **b.haig@psyc.canterbury.ac.nz** **www.psyc.canterbury.ac.nz**

Abstract: Chow's endorsement of a limited role for null hypothesis significance testing is a needed corrective of research malpractice, but his decision to place this procedure in a hypothetico-deductive framework of Popperian cast is unwise. Various failures of this version of the hypothetico-deductive method have negative implications for Chow's treatment of significance testing, meta-analysis, and theory evaluation.

Believing that the various criticisms of null hypothesis significance testing (NHST) have been made without a coherent framework for understanding research procedures, Chow (1996; 1998b) proceeds to embed NHST in the context of theory corroboration experiments. His adoption of Popperian hypothetico-deductive method as his framework is unwise, however. Some of the various defects of this popular account of method mar aspects of his treatment of NHST, meta-analysis, and theory corroboration. In this commentary I suggest that: (1) NHST can also be used to calibrate scientific instruments; (2) the basic goal of meta-analysis is to help us detect empirical phenomena; and (3) abduction and inference to the best explanation are important elements of theory evaluation in science.

Calibration and significance testing. With its predilection for "Fisherian" experiments, psychological research methodology tends to ignore a range of important strategies that provide justification for the belief in experimental results. One such strategy is calibration, which involves using a substitute signal to standardize a measuring instrument (Franklin 1997). By focusing on the place of NHST in theory corroboration, Chow fails to consider that significance testing can also serve as a useful test for the calibration of scientific instruments. Unfortunately, because calibration itself is neglected in orthodox treatments of psychological experimentation, this use of significance testing to check on calibration is widely ignored in the psychological literature.

In science, instruments must be calibrated before they can be used in a trustworthy manner; because instruments tend to go out of calibration, they may need to be recalibrated. Of course, even with properly calibrated instruments some random error is to be expected. To test whether measured values from an instrument represent chance fluctuations or signal a loss of calibration, we can use a test of significance (Baird 1992). The normal curve is widely used as a model of chance fluctuations or errors of measurement. In this context, errors can be thought to result from numerous small, independent disturbances, such as slight variations in the mechanical or electrical components of the measuring instrument. At times such disturbances will produce measurement readings that are too large, on other occasions, readings that are too small. Where the probability of such tendencies is p, a binomial distribution represents the small disturbances pushing either way of the true value. Because the normal distribution approximates such binomial probabilities well, we can use NHST as a check on calibration.

Phenomena detection and meta-analysis. A little recognized limitation of the hypothetico-deductive method stems from its acceptance of the widespread view that scientific theories explain and predict facts about observed *data.* Strictly speaking, however, this is a misleading portrayal of the nature of science, for it is *phenomena,* not data, that typically constitute the proper objects of scientific explanation (Woodward 1989). Phenomena are relatively stable, recurrent general features of the world we seek to explain and predict. Data, by contrast, lack the stability and generality of phenomena, being idiosyncratic to particular investigative contexts. Whereas the importance of data lies in the fact that they serve as evidence for the phenomena under investigation, phenomena provide the evidence for theories, and because of their generality and stability, become the appropriate foci of scientific explanation. Because of his commitment to the hypothetico-deductive method, Chow fails to see that phenomena detection is an important goal of scientific discovery.

Unfortunately, this failure to attend to phenomena detection unduly limits Chow's examination of meta-analysis. Chow maintains that efforts to show the superiority of meta-analysis over NHST are misplaced. To this end, he plausibly argues that meta-analysis cannot function as a theory corroboration procedure. Chow fails to appreciate, however, that in calculating effect sizes across primary studies in a common domain, meta-analysis helps us detect "ubiquitous positive effects" (Schmidt 1992). In other words, meta-analysis is concerned with the detection of phenomena that will comprise the basic facts that stand in need of scientific explanation. Hunter (1998) is right that the objective of meta-analysis is to discover the facts. It is in this role that I think meta-analysis performs its most useful work for science. By enabling us to use statistical methods to ascertain the existence of robust empirical regularities, meta-analysis can usefully be viewed as the statistical analogue of experimental replication.

Abduction and theory generation. In defending the place of NHST in a hypothetico-deductive framework, Chow criticizes meta-analysts such as Schmidt (1992) for claiming that empirical research is carried out in psychology prior to the discovery of theory. He echoes Popper's twin beliefs that there is no method for moving from phenomena to explanatory theory and that empirical research is conducted to justify conjectural theories. However, to deny method a place in theory generation goes against a good deal of scientific practice, wherein scientists endeavor to explain data patterns or empirical phenomena of interest by postulating underlying causal mechanisms through an abductive or theoretical reasoning process (Josephson & Josephson 1994). Abductive inference is concerned with the initial conceiving of an idea or hypothesis. It is a form of explanatory inference and often involves reasoning from presumed effects to underlying cause. In psychology, exploratory factor analysis provides a good example of a theory generation method that enables the researcher to reason abductively from correlational data to underlying common causes.

Theories arrived at via abduction are not unconstrained speculations but educated guesses constrained by relevant knowledge in the domains under investigation. Assessments of the initial plausibility of such theories basically involve ascertaining whether they are the products of sound abductive reasoning. In arriving at such assessments we use a generative form of justification. Generative methods, such as exploratory factor analysis, reason from warranted premises to an acceptance of the knowledge claims in question (Nickels 1987). These methods complement the consequentialist methods, such as hypothetico-deductivism, which reason from the knowledge claims in question to their testable consequences.

Contra hypothetico-deductive dogma, there is a logic or rationality to theory generation provided by abductive reasoning; and there is a generative form of justification concerned with assessing the prospective worth of our newly generated theories. An acknowledgment of these two points by Chow would have enabled him to appreciate the role of meta-analysis in the detection of empirical phenomena and, in turn, the fact that phenomena provide the natural press for generating explanatory theories.

Inference to the best explanation and theory evaluation. Throughout his book, Chow (1996) makes it clear that NHST services the corroboration of explanatory theories. In his Response, Chow (1998b) discusses the merits of good theories and stresses the importance of their explanatory nature. He does not consider, however, the important fact that such theories will have explicitly explanatory values that are relevant to an assessment of their goodness. This is not surprising, for in working within a hypothetico-deductive framework, Chow naturally adopts the widely held view that the mark of a good explanatory theory is its predictive success.

In scientific methodology, the place of prediction in ascertaining the empirical adequacy of scientific theories has been emphasized at the expense of their explanatory value. Scientists themselves often judge the worth of their theories with explanatory criteria in mind (Thagard 1992). Such a view of theory evaluation employs a style of reasoning known as *inference to the best explanation.* Inference to the best explanation is a form of abductive inference that leads to the acceptance of the best of competing theories that endeavor to explain the relevant phenomena.

Critics have sometimes pointed out that formulations of inference to the best explanation are too vague, either because they fail to tell us how such inferences are to be made or because they object to a particular formulation of this type of inference, such as that noted by Erwin (1998) in his questioning of Chow's reconstruction of theory corroboration. Recently, however, significant progress has been made in understanding how one might successfully employ inference to the best explanation to evaluate the worth of competing theories. Thagard (1992) has developed an important account of theory evaluation that takes inference to the best explanation to be centrally concerned with establishing explanatory coherence. Chow (1998b) asserts that "[t]he minimal criterion for accepting an explanatory hypothesis is that it should be consistent with the phenomenon to be explained" (p. 233). Presumably, for Chow this will be the *logical* coherence of deductive consistency wrought from Popperian hypothetico-deductive method. For Thagard, however, the consistency comes in the form of *explanatory* coherence where the propositions hold together because of their explanatory relations.

According to Thagard's model, the explanatory coherence of a theory is determined by considering three criteria: explanatory breadth (or consilience), simplicity, and analogy. Explanatory breadth, which is the most important criterion, captures the idea that a theory is more explanatorily coherent than its rivals if it explains a greater range of facts. The notion of simplicity is also important for theory choice, and is captured by the idea that preference should be given to theories that make fewer special assumptions. With the third criterion – analogy – explanations are judged more coherent if they are supported by analogy to theories that scientists already find credible. Using the theory of explanatory coherence to make inferences to the best explanation will often be a better way to justify explanatory theories than using the hypothetico-deductive method to establish their predictive success. Whenever this is so, NHST will be irrelevant.

In conclusion, Chow's discussion of NHST is distorted by his use of the restrictive Popperian hypothetico-deductive account of scientific method. The broader conception of scientific method assumed in this commentary suggests a different make up of the psychological researcher's armamentarium from that suggested by Chow: NHST should receive very limited use, but in a theory-testing context that is more defensible than that provided by the simplistic hypothetico-deductive method; it can also be used in measurement contexts where calibration checks are needed; meta-analysis should be used primarily to provide the statistical replication needed to detect phenomena from the numerous primary studies in a given domain; and, codified abductive procedures should be employed as methods of theory construction, both for the generation of explanatory theories and for their evaluation in terms of explanatory coherence.

Does the finding of statistical significance justify the rejection of the null hypothesis?

David Sohn

Department of Psychology, University of North Carolina at Charlotte, Charlotte, NC 28223–0001. **dsohn@email.uncc.edu**

Abstract: The soundness of Chow's (1996; 1998a) argument depends on the soundness of his assertion that statistical significance may be understood to signify that chance may be *excluded* as the reason for results. The examples and arguments provided here show that statistical significance signifies no such thing.

Chow defends the significance test by making what he calls a limited claim for its usefulness: "It provides a rational basis for excluding chance influences as an explanation of data" (Chow 1998a, p. 176). Because the null hypothesis, herein called the hypothesis of chance, is mutually exclusive with the hypothesis of nonchance, its disaffirmation justifies the affirmation of the hypothesis that nonchance influences are responsible for the result. The occasion for excluding chance is the finding of statistical significance: "A statistically significant result will be correctly interpreted to mean only that an explanation of the data in terms of chance influences can be excluded with the level of strictness stipulated by the significance level (viz., α)" (Chow 1998a, p. 176). Not one of the writers of the 37 commentaries expresses disagreement with this view of the function of the significance test. Has Chow put his finger on something that is at last inarguably a function the test *is* able to perform?

The answer is "no," because there is a crucial element missing in Chow's argument. This is the demonstration that a statistically significant result *justifies* the rejection of the chance hypothesis. Chow provides no real argument, explanation, or demonstration to support his contention, choosing instead to proclaim the reasonableness of the view that a test statistic with a small enough associated p justifies the rejection of the chance hypothesis: "It is reasonable to ignore an outcome if its probability of occurrence is small enough. The question is how small the probability of an infrequent event should be . . . before it becomes reasonable to reject chance as an explanation" (Chow 1996, p. 36).

Chow is saying that if an outcome were to occur, let us say, one in a million times under the hypothesis of chance, it would be *reasonable* to reject the hypothesis of chance. But how can one tell from the result alone that the outcome is not that one-in-a-million occurrence? And by what reasoning does one arrive at the view that it is a good "bet" that the chance hypothesis is incorrect when one has an outcome that is improbable under the terms of the chance hypothesis?

In fact, the reasonableness of rejecting the chance hypothesis on the basis of a statistically significant result has been in question since the beginning of significance testing. Consider this attack on the logic of the practice by Berkson:

> The argument seems to be illogical. Consider it in symbolic form. It says, "If *A* is true, *B* will happen sometimes; therefore if *B* has been found to happen, *A* can be considered disproved." There is no logical warrant for considering an event known to occur in a given hypothesis, even if infrequently, as disproving the hypothesis. (1942, p. 326)

Much later, I presented arguments to show that statistical significance is not an augury of the truth of the nonchance hypothesis (Sohn 1993). What follows are reasons and explanations that show why the finding of statistical significance does not signify the falsity of the chance hypothesis.

Consider the case of 1,000 independent two-condition experiments. A directional t-test, using an α of .05, is performed in each case. Suppose that in *all* cases, chance is responsible for the outcome. One can expect, then, that around 5%, or 50, of the findings will be significant, leading, in each case, to the rejection of the chance hypothesis. All of these rejections will be Type I errors; there will be *no* correct exclusions of the chance hypothesis. The 950 acceptances of the chance hypothesis will be correct, but these are *inclusions,* not exclusions, of chance as the explanation.

Suppose the opposite scenario, the nonchance hypothesis, is correct in all cases. Now every rejection of chance is correct. If it can be determined how many or what percentage of the decisions will be rejections, it will be possible to evaluate the performance of the significance test in excluding chance. There is a problem. The likelihood of a decision being one to reject depends on the size of the nonchance effect. If it is large, the decision to reject is more likely than if the effect is small. In the case of huge effects, the decision to reject will be reached every time. In such a circumstance, all decisions will be rejections of chance, and correct ones at that. On the other hand, in the case of minuscule effects, the percentage of decisions to reject may be scarcely greater than α. In this case, the score card could be around 50 correct decisions to reject versus 950 incorrect failures to reject the chance hypothesis. But because the effect sizes are unknown, it is simply a guess what the performance of the test is in rejecting the chance hypothesis on the basis of statistical significance.

For scenarios involving a mix of both true chance and nonchance hypotheses, the principles are the same as for the two cases considered. There will be no correct exclusions of chance when the chance hypothesis is true, and the success rate will be indeterminate in the case of true nonchance hypotheses. Because there is no knowledge of the frequency of truth of the chance and nonchance hypotheses or of the size of nonchance effects in the case of true nonchance hypotheses, there can be no way to assess the test's performance in excluding chance. (If such knowledge were available, there would be no need for significance testing.)

Thus in no possible scenarios is there a basis for the claim that statistical significance implies that chance may be excluded as the explanation of data. Differently said, there appears to be no reason to believe that it is a good bet that the chance hypothesis is false when there is a finding of statistical significance.

Author's Response

The Popperian framework, statistical significance, and rejection of chance

Siu L. Chow

Department of Psychology, University of Regina, Regina, Saskatchewan, Canada S4S 0A2. **siu.chow@uregina.ca** **uregina.ca/~chowsl/**

Abstract: That **Haig** and **Sohn** find the hypothetico-deductive approach wanting in different ways shows that multiple conditional syllogisms are being used in different stages of theory corroboration in the Popperian approach. The issues raised in the two commentaries assume a different complexion when certain distinctions are made.

Separate conditional syllogisms at different levels of abstraction are being used to justify the rejection of chance in significance tests and to corroborate theories. **Haig**'s concerns are met with a discussion of the nature of psychometric instruments, the incommensurability problem of meta-analysis, and the circularity of adductive conclusions. Simulation data are used to answer **Sohn**'s critiques by showing that (1) the alpha level is not meant to be applied to a set of t-tests, and (2) statistical significance is dependent on neither effect size nor sample size.

It is important to calibrate an instrument if it is used to obtain exact measurements (e.g., a clock). Psychologists do not use significance tests for calibration purposes because psychometric instruments provide relative, not absolute, measurements. For example, a WISC-R score of 115 indicates, not how intelligent an individual is, but that the individual is better than 84.13% of the norm group. The acceptability of a psychometric instrument depends on its validity, not the sort of precision monitored by calibration.

The 12 studies described in Chow's (1996) Table 5.5 belong to the same domain. To conduct a meta-analysis on them is to obtain the average of the diverse effects of the qualitatively different independent variables. The result is conceptually anomalous because it is not theoretically meaningful to mix apples and oranges. As no valid conclusion can be drawn from meta-analysis, it cannot be used to discover new phenomena.

Haig's concerns about theory discovery and the interrelationships among phenomenon, theory, and evidential data have been anticipated in sections 3.2.1 (pp. 46–47) and 3.7 (pp. 63–64) in Chow (1996). Given the "phenomenon → hypothesis → evidential data" sequence (Chow 1996, pp. 46 and 63), the theory is necessarily an *ad hoc* postulation *vis-à-vis* the to-be-explained phenomenon. It is circular for Haig to assert that "phenomena provide the evidence for theories." Hence, theories obtained by adduction, like theories established by any other means, have to be corroborated. A series of three embedding conditional syllogisms is used when corroborating theories, including adductively established ones (Chow 1996, Table 4.2, p. 70).

The null hypothesis (H_0) is used in significance tests as the antecedent and the consequent of two conditional propositions (Chow 1996, p. 32) as follows:

> [Proposition 1]: *If* the research manipulation is not efficacious (i.e., only chance influences are assumed), *then* H_0.
> [Proposition 2]: *If* H_0, *then* the mean difference of the sampling distribution of differences is zero.

This practice of emphasizing that H_0 is an implication of the change hypothesis, not the chance hypothesis itself, will henceforth be called the formal approach. **Sohn**'s treatment of H_0 as the chance hypothesis is acceptable as a casual way to express the transitive relationship between Propositions 1 and 2. Subsequently, it is called the vernacular approach.

There are important differences between the formal and vernacular stances. For example, Proposition 2 is true only if H_0 is the zero-null (i.e., H_0; $u_1 = u_2$). If H_0 is a point-null (e.g., u_1–u_2 = 5), Proposition 2 is replaced by Proposition 3:

> [Proposition 3]: *If* H_0: u_1–u_2 = 5, *then* the mean difference of the sampling distribution of differences is 5.

Not making the distinction between the formal and vernacular approaches is responsible for some of the issues raised by **Sohn.**

If H_0 were the chance hypothesis, "significant" and "rejecting the chance hypothesis" become synonymous characterizations. The question about justification becomes mute because two synonymous expressions do not (and cannot) have a justificatory relationship. On the other hand, excluding chance as an explanation by rejecting H_0 in the case of Proposition 1 is warranted by *modus tollens* (Chow 1996, pp. 50–52).

Sohn questions the justificatory function of *modus tollens* because of Berkson's (1942) conditional syllogism. If one were to follow Berkson's example, *A* represents the chance hypothesis, and *B* stands for H_0. This is not possible when H_0 is the chance hypothesis (viz., Sohn's contention), however, because a concept does not imply itself. Moreover, the "sometimes" qualifier makes ambiguous Berkson's major premise. Furthermore, there is a confusion in Sohn's appeal to Berkson.

> [Proposition 4]: Of all possible differences between two sample means, 5% produce a *t*-value equal to, or smaller than, the critical *t* value.
>
> [Proposition 5]: Some differences between two sample means produce a *t* value equal to, or smaller than, the critical *t* value.

Proposition 4 is a definite statement about a probabilistic phenomenon that can be tested. The ambiguity of Berkson's (1942) minor premise (like Proposition 5) precludes it from being used as a criterion for making the statistical decision. Subscribing to Berkson's reasoning betrays a confusion between adopting a well-defined probabilistic statement and using a vague proposition.

Sohn finds the reasonableness of the formal approach wanting because one can never be certain that H_0 is false when one rejects it. This objection would be unassailable if absolute certainty were the prerequisite for reasonableness. Be that as it may, the inevitable uncertainty in question does not invalidate the formal approach.

The Type I error becomes a concern when there are reservations about the statistical significance of the result of a specific experiment. This is an occasion for checking the correctness of the experimental hypothesis, the presence of a confounding variable, or the appropriateness of the experimental design, task or procedure. That is, instead of disputing the validity, usefulness or importance of significance tests, the inevitable uncertainty serves to ensure conceptual or methodological rigor.

Sohn's two scenarios set in high relief a common misunderstanding about significance tests. Specifically, it is said in the first scenario (Sohn's para. 5) that, given $\alpha = .05$, the results of around 50 of 1,000 separate *t*-tests will be significant by chance when the zero-null hypothesis is true. This statement is as incorrect as saying that there will be *n* heads and *n* tails in 2*n* identical tosses of a fair coin. What a fair coin implies is that 50% of an infinite number of identical tosses result in heads. It does not follow that half of any exact number of identical tosses will result in heads. Consider the first scenario more closely with reference to Table R1.

Underlying the *t*-test are two statistical populations specified by the two levels of the independent variable (Winer 1962). Shown in Panel 1 of Table R1 is the composition of two such populations. Their means are shown in Panel 2 (viz., $u_1 = u_2 = 4.812$) and they have the same standard deviation (viz., $\sigma_1 = \sigma_2 = .894$). The following steps were carried out:

(a) Selected with replacement a random sample of n_1 units from Population 1 and another random sample of n_2 from Population 2, and $n_1 = n_2$.

(b) Ascertained the difference between the two sample means, as well as the standard error of the difference.

(c) Calculated the *t* ratio.

(d) Returned the two sets of *n* units to their respective populations.

(e) Repeated steps (a) through (d) 5,000 times.

(f) Steps (a) through (e) were repeated with $n_1 = n_2 =$ 5, 75, 750, and 1,000.

Given any sample size, there are 5,000 differences at the end of the exercise. When they are tabulated in the form of a frequency distribution, the result is an empirical approximation to the random sampling distribution of the differences between two sample means. It is only an approximation because, in theory, step (e) should consist of an infinite number of times.

The 5,000 *t*-values obtained in step (c) represent the result of standardizing the 5,000 differences in terms of their respective standard errors of differences. Shown in Column 2A of Table R1 are the numbers of empirically determined *t*-values that fall within the ranges identified in the corresponding row. For example, 104 *t*-values fall between −1.90 and −1.701. This simulation exercise is to make explicit four points:

(1) The probability foundation of the *t*-test is the sampling distribution of differences.

(2) A different sampling distribution of differences is used when the sample size changes (see columns 2A through 2D of Table R1).

(3) The expression "$\alpha = .05$" means that 5% of an infinite number of differences between 2 means give *t*-values that are as extreme as, or more extreme than, 1.86 (or −1.86 as the case may be) for the 1-tailed test with df = 8.

(4) It does not follow from (3) that 5% of any 1,000 differences would be as extreme as, or more extreme than, the critical *t* value.

To recapitulate (1), every application of the *t*-test evokes the appropriate sampling distribution of differences. Hence, the same sampling distribution is evoked 1,000 times in **Sohn**'s first scenario if the 2 statistical populations (as well as n_1 and n_2) remain the same throughout. The 50–950 split of the 1,000 experiments envisioned by Sohn has nothing to do with the alpha level for the reasons stated in (3) and (4).

Given that testing a point-null hypothesis is no different from testing a zero-null hypothesis (Kirk 1984), the outcomes of significance tests should be independent of the ex-

Table R1: The composition of two identical statistical populations used in simulation (Panel 1) and four distributions of 5,000 t-*ratios when the zero-null is true (Panels 2A through 2D)*

Panel 1	Score	1	2	3	4	5	6	7	8	Total
	Frequency	1	12	36	412	669	128	65	5	1328

Panel 2	$u_1 = u_2 = 4.182$; $N_1 = N_2 = 1328$; $\sigma_1 = \sigma_2 = .894$			
	2A	2B	2C	2D
Range of *t* ratio's	$u_1 - u_2 = 0$ $n_1 = n_2 = 5$	$u_1 - u_2 = 0$ $n_1 = n_2 = 75$	$u_1 - u_2 = 0$ $n_1 = n_2 = 750$	$u_1 - u_2 = 0$ $n_1 = n_2 = 1000$
	Frequency	Frequency	Frequency	Frequency
≤−2.901	44	12	4	13
−2.90–−2.701	8	9	11	7
−2.70–−2.501	41	12	19	14
−2.50–−2.301	44	30	22	15
−2.30–−2.101	58	37	45	38
−2.10–−1.901	11	67	53	56
−1.90–−1.701	104	79	85	75
−1.70–−1.501	76	117	92	124
−1.50–−1.301	192	172	160	137
−1.30–−1.101	175	181	187	196
−1.10–−.901	273	237	228	231
−.900–−.701	227	279	289	282
−.700–−.501	362	307	335	338
−.500–−.301	403	340	398	371
−.300–−.101	111	342	399	368
−.100–.099	723	494	414	382
.100–.299	104	358	402	424
.300–.499	413	367	321	385
.500–.699	384	350	319	345
.700–.899	237	293	283	293
.900–1.099	260	234	269	226
1.10–1.299	170	204	201	200
1.30–1.499	173	149	138	149
1.50–1.699	121	118	102	105
1.70–1.899	112	69	73	76
1.90–2.099	10	55	53	55
2.10–2.299	71	32	34	36
2.30–2.499	39	22	32	26
2.50–2.699	27	14	18	12
2.70–2.899	8	10	6	8
2.90–3.099	1	7	3	6
≥3.100	18	3	5	7
	$s_{\bar{x}_1 - \bar{x}_2} = .560$	$s_{\bar{x}_1 - \bar{x}_2} = .147$	$s_{\bar{x}_1 - \bar{x}_2} = .046$	$s_{\bar{x}_1 - \bar{x}_2} = .04$
Mean *t*-ratio	−.009	−.005	−.006	.006
Expected *t*-ratio	0	0	0	0

pected effect size (Chow 1996, pp. 132–34; 1998a, pp. 184–85). This contradicts the second scenario described in **Sohn**'s paragraph six, which is an echo of the power-analytic "significance-effect size dependence" assertion that the outcomes of significance tests depend on effect size (Cohen 1987). This point is amplified below.

The entries in Table R2 were also obtained with steps (a) through (f), except that the mean of the second statistical population is larger than that of the first one by 0.5 of the standard deviation of the first statistical population (viz., $u_1 = 4.812$; $u_2 = 5.262$; Panel 2). If the "significance-effect size dependence" thesis were correct, the mean *t*-ratio should differ from zero. There is no support for the "significance-effect size dependence" thesis because none of the four mean *t*-ratios differs from 0 (viz., .028, .013, .012, and .008).

Sohn's second scenario also echoes another power-analytic assertion, namely, that larger sample sizes increase sta-

Table R2: The composition of the control statistical population used in simulation (Panel 1) and 4 distributions of 5,000 t*-ratios when the point-null is true (Panels 2A through 2D)*

Panel 1	Score	1	2	3	4	5	6	7	8	Total
	Frequency	1	12	36	412	669	128	65	5	1328

Panel 2	$u_1 = 4.182; u_2 = 5.262; N_1 = N_2 = 1328; \sigma_1 = \sigma_2 = .894$			
	2A	2B	2C	2D
Range of *t* ratio's	$u_1 - u_2 = 0$ $n_1 = n_2 = 5$	$u_1 - u_2 = 0$ $n_1 = n_2 = 75$	$u_1 - u_2 = 0$ $n_1 = n_2 = 750$	$u_1 - u_2 = 0$ $n_1 = n_2 = 1000$
	Frequency	Frequency	Frequency	Frequency
≤−2.901	19	6	12	3
−2.90−−2.701	9	12	15	5
−2.70−−2.501	37	17	11	12
−2.50−−2.301	33	20	10	10
−2.30−−2.101	67	39	30	36
−2.10−−1.901	13	65	43	51
−1.90−−1.701	105	80	90	100
−1.70−−1.501	62	114	125	119
−1.50−−1.301	172	147	139	173
−1.30−−1.101	171	179	180	180
−1.10−−.901	237	223	237	258
−.900−−.701	224	261	302	327
−.700−−.501	385	299	339	331
−.500−−.301	392	322	356	332
−.300−−.101	116	404	395	337
−.100−.099	850	497	406	405
.100−.299	100	377	365	360
.300−.499	388	359	380	357
.500−.699	379	353	316	332
.700−.899	251	313	292	315
.900−1.099	243	246	259	254
1.10−1.299	153	199	204	197
1.30−1.499	162	142	150	172
1.50−1.699	136	114	115	108
1.70−1.899	109	79	97	82
1.90−2.099	9	51	49	49
2.10−2.299	81	31	30	43
2.30−2.499	25	17	20	22
2.50−2.699	32	13	11	10
2.70−2.899	13	6	10	9
2.90−3.099	1	7	8	6
≥3.100	26	8	4	5
	$s_{\bar{x}_1-\bar{x}_2} = .589$	$s_{\bar{x}_1-\bar{x}_2} = .145$	$s_{\bar{x}_1-\bar{x}_2} = .046$	$s_{\bar{x}_1-\bar{x}_2} = .04$
Mean *t*-ratio	.028	.013	.012	.008
Expected *t*-ratio	0	0	0	0

tistical power, thereby making it easier to obtain statistical significance. This "significance-sample size dependence" thesis is questioned by the four χ^2 tests reported in the two panels of Table R3, as may be seen from the italicized and boldface entries in the two panels.

Use Panel B of Table R3 as an illustration. Each of the 5,000 *t*-values in Column 2A of Table R2 was classified as "Significant" or "Not significant." For example, there are 456 and 4544 *t*-values in the "Significant" and "Not significant" categories, respectively, when $n_1 = n_2 = 5$. The same process was repeated with the entries from each of the other columns of Table R2 (i.e., for sample sizes of 75, 750, and 1,000). The result is the eight boldface entries in Panel B of Table R3. They make up the two-way χ^2 test for the independence of statistical significance (columns) and sample size (rows). As the $\chi^2 = 2.64$ ($df = 3$) is not significant, there is no reason to reject the independence in question. That is, there is no support for the view that statistical sig-

Table R3: The number of empirically determined t*-ratios tabulated in Tables 1 (Panel A; zero-null) and 2 (Panel B; point-null) that exceed the critical value of the* t*-ratio (significant) and do not exceed the critical value (nonsignificant) at $\alpha = 0.05$, 2-tailed*

		df	$n_1 = n_2$	Critical *t*	Significant	Not significant	χ^2 (*df* = 3)
		8	5	≤−1.86 or ≥1.86	*465*	*4535*	
A	a = 0.05	148	75	≤−1.65 or ≥1.65	*490*	*4510*	
	(2-tailed)	1498	750	≤−1.645 or ≥1.645	*514*	*4486*	2.923
	(Zero-null)	1998	1000	≤−1.645 or ≥1.645	*501*	*4499*	
		8	5	≤−1.86 or ≥1.86	**456**	**4544**	
B	a = 0.05	148	75	≤−1.65 or ≥1.65	**484**	**4516**	
	(2-tailed)	1498	750	≤−1.645 or ≥1.645	**493**	**4507**	2.64
	(Point-null)	1998	1000	≤−1.645 or ≥1.645	**501**	**4499**	

nificance is a function of sample size in the case of the point-null.

The procedure just described was also carried out with the entries of Table R1. The result is the eight italicized entries in Panel A of Table R3. The χ^2 of 2.93 (*df* = 3) is also not significant. Hence, there is also no support for the "significance-sample size dependence" thesis in the case of the zero-null.

To conclude, it is necessary to distinguish between (a) phenomenon and evidential data, and (b) the chance hypothesis and H_0. The inevitable possibility of committing the Type I error does not invalidate the formal approach to significance tests. The exclusion of the chance explanation by rejecting H_0 is warranted by *modus tollens.* Although the critical *t*-value defined by the alpha level serves as the decision criterion in every *t*-test, it has nothing to do with a collection of separate *t*-tests as a set.

References

Letters "a" and "r" appearing before authors' initials refer to target article and response, respectively

Baird, D. (1992) *Inductive logic: Probability and statistics.* Prentice Hall. [BDH]

Berkson, J. (1942) Tests of significance considered as evidence. *Journal of the American Statistical Association* 37:325–35. [DS]

Chow, S. L. (1996) *Statistical significance: Rationale, validity, and utility.* Sage. [rSLC, BDH, DS]

(1998a) Précis of *Statistical significance: Rationale, validity, and utility. Behavioral and Brain Sciences* 21:169–239. [rSLC, DS]

(1998b) The null-hypothesis significance-test procedure is still warranted. *Behavioral and Brain Sciences* 21:228–38. [BDH]

Cohen, J. (1987) *Statistical power analysis for the behavioral sciences* (revised edition). Academic Press. [rSLC]

Erwin, E. (1998) The logic of null hypothesis testing. *Behavioral and Brain Sciences* 21:197–98. [BDH]

Franklin, A. (1997) Calibration. *Perspectives on Science* 5:31–80. [BDH]

Hunter, J. E. (1998) Testing significance testing: A flawed defense. *Behavioral and Brain Sciences* 21:204. [BDH]

Josephson, J. R. & Josephson, S. G., eds. (1994) *Abductive inference.* Cambridge University Press. [BDH]

Kirk, R. E. (1984) *Basic statistics,* 2nd edition. Brooks/Cole. [rSLC]

Nickles, T. (1987) Methodology, heuristics and rationality. In: *Rational changes in science,* ed. J. C. Pitt & M. Pera. Reidel. [BDH]

Schmidt, F. L. (1992) What do data really mean? Research findings, meta-analysis, and cumulative knowledge in psychology. *American Psychologist* 47:1173–81. [BDH]

Sohn, D. (1993) Psychology of the scientist: LXVI. The idiot savants have taken over the psychology labs! Or why in science the rejection of the null hypothesis as the basis for affirming the research hypothesis is unwarranted. *Psychological Reports* 73:1167–75. [DS]

Thagard, P. (1992) *Conceptual revolutions.* Princeton University Press. [BDH]

Winer, B. J. (1962) *Statistical principles in experimental design.* McGraw-Hill. [rSLC]

Woodward, J. (1989) Data and phenomena. *Synthese* 79:393–472. [BDH]