Psychological Review
2010, Vol. 117, No. 1, 300–308


0033-295X/10/$12.00 DOI: 10.1037/a0018047

# More on Grandmother Cells and the Biological Implausibility of PDP Models of Cognition: A Reply to Plaut and McClelland (2010) and Quian Quiroga and Kreiman (2010)

Jeffrey S. Bowers
University of Bristol

Plaut and McClelland (2010) and Quian Quiroga and Kreiman (2010) both challenged my characterization of localist and distributed representations. They also challenged the biological plausibility of grandmother cells on conceptual and empirical grounds. This reply addresses these issues in turn. The premise of my argument is that grandmother cells in neuroscience are the equivalent of localist representations in psychology. When defined in this way, grandmother cells are biologically plausible, given the neuroscience to date. By contrast, the neurophysiology is shown to be inconsistent with the distributed representations often learned in existing parallel distributed processing (PDP) models, and it poses a challenge to PDP theories more generally.

*Keywords:* grandmother cells, localist representation, distributed representation, PDP models, connectionism

Plaut and McClelland (2010) emphasized that parallel distributed processing (PDP) models are not intended to capture all the neurophysiology that underpins cognition. Rather, the models are designed to account for various behavioral phenomena and should be evaluated in these terms. Nevertheless, these authors took the successes of these models at the behavioral level as evidence that the PDP approach captures something fundamental about the brain. That is, they took the successes as evidence that the brain relies on distributed representations when identifying words, objects, and faces.

The problem with this analysis is that PDP models are rarely judged on the basis of their behavioral successes alone. For example, Seidenberg and Plaut (2006) contrasted PDP and localist (grandmother cell) models of word naming and acknowledged that localist models often provide a better account of the empirical findings. They nevertheless endorse the PDP approach because distributed representations are considered more biologically plausible. This claim regarding the brain is the received wisdom in the cognitive sciences, and it strongly impacts how the empirical successes of localist and distributed models are evaluated. If models were only assessed at the behavioral level, then localist models would be more widely supported. Indeed, adopting the logic of Plaut and McClelland (2010), the successes of localist models at a behavioral level could be taken as evidence that the brain relies on grandmother cells.

In Bowers (2009), I considered the claim that PDP models capture something fundamental about the workings of the brain, by reviewing the large literature on single cell neurophysiology. The conclusions I reached are very much at odds with the PDP approach. That is, I argued that the neuroscience falsifies many existing PDP models but is consistent with localist ones. This is not to say that the neuroscience uniquely supports a grandmother theory of neural coding. My claim is that only some versions of distributed coding are falsified by single-cell recording studies and that additional empirical and computational research is required to test the relative merits of grandmother and alternative distributed coding schemes.

Plaut and McClelland (2010) and Quian Quiroga and Kreiman (2010) disagreed with my analysis in a number of respects. First, they challenged my characterization of localist and distributed representations, and Plaut and McClelland challenged my characterization of the PDP approach as well. Second, Plaut and McClelland challenged the biological plausibility of grandmother cells on the basis of on computational considerations (e.g., local coding schemes cannot support the identification of specific instances of objects, such as Bowers's breakfast toast). Finally, Quian Quiroga and Kreiman challenged the plausibility of grandmother cells on the basis of on neurophysiological data. I respond to these issues in turn.

## Definitions of Localist Versus Distributed Coding

At first blush, the distinction between localist and distributed coding schemes would seem to be straightforward, but as made clear in the above two commentaries, there are fundamental points of disagreement (also see Page, 2000, and the associated commentaries). This is not simply a terminological issue, as the different use and understanding of the terms lead to different interpretations of data and contrasting theoretical conclusions.

## Grandmother Cells

The premise of the target article is that grandmother cell theories in neuroscience are the equivalent of localist theories in psychology. The core claim in both domains is that perception is organized

I thank Colin Davis, David Plaut, and Rodrigo Quian Quiroga for reading and commenting on drafts of this article.

Correspondence concerning this article should be addressed to Jeffery S. Bowers, Department of Experimental Psychology, University of Bristol, 12A Prior Road, Clifton, Bristol BS8-ITU, England. E-mail: j.bowers@bris.ac.uk

in a hierarchy of processing steps, with neurons (processing units) at higher levels of the hierarchy pooling information from lower levels, such that each subsequent stage codes for something more complex. At the top of the hierarchy are single neurons (units) that code for words, objects, and faces. These theories are a logical extension of the pioneering work of Hubel and Wiesel (1968), who found that low-level vision is organized in just this way, with complex cells in primary visual cortex (V1) pooling inputs from multiple simple cells in order to code for more complex stimuli. In fact, Hubel (1995) considered the implications of this hierarchal organization within early vision and raised the question of whether the same design principles apply throughout, with grandmother cells at the top. He rejected this hypothesis as implausible.

Although Hubel (1995) was skeptical that this hierarchy extends to high-level vision, localist models adopt exactly this logic. For example, the interactive activation (IA) model includes single units at the earliest level that respond to line segments (much like simple cells), which pool their outputs to localist letter representations, which in turn pool their outputs onto localist word representations (McClelland & Rumelhart, 1981). Word identification is achieved when a single word unit is activated beyond some threshold. My basic claim is that the hierarchical structure of the IA model and its use of localist word representations is biologically plausible—more so than are PDP models of word identification that do not have either of these properties.

This characterization of grandmother cell theories is challenged by theorists who define distributed representations much more broadly. For example, Figure 1 is an adaptation of a figure taken from Poggio and Bizzi (2004), who described a model of face identification inspired from single cell recording studies. The model includes simple and complex cells in V1, units in infratemporal (IT) cortex tuned to specific familiar faces at a given viewing angle, and units in anterior IT (AIT) tuned to familiar faces independent of orientation. The model is described as an extension of the original Hubel and Wiesel model (1968). Despite the fact that a familiar face can be identified when a single unit is activated beyond some threshold, they characterize this model as distributed because an image of a familiar face activates more than one face unit (as a function of similarity). Although the coactive face units play no role representing familiar faces, Poggio and Bizzi argued that the coactivation of face units is required for generalization to novel faces and explicitly ruled out grandmother cell theories on this basis.

The problem with this argument is that it renders the concept of grandmother (or localist) representation trivial—it is a hypothesis that no one would defend. As discussed in the target article, all localist models in psychology have the property that multiple localist units are coactivated by familiar inputs, and these coactive representations can impact the identification of familiar (and unfamiliar) stimuli. When evaluating the plausibility of grandmother cell theories, the critical question is not whether more than one unit is activated in response to an input but rather whether the identification of a familiar thing (word, object, face) can be inferred on the basis of a single unit being active beyond some threshold. This is how the Poggio and Bizzi (2004) model works. It is a grandmother cell theory, exactly the sort theory that Hubel (1995) rejected.

Plaut and McClelland (2010) adopt an even broader definition of distributed coding. They argued that words in the IA model are

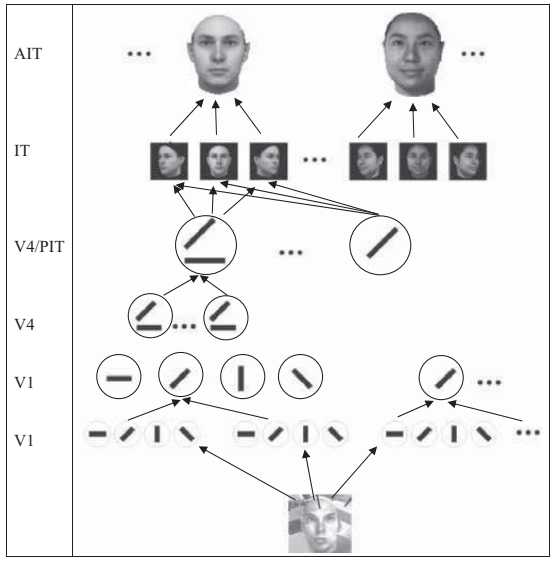


*Figure 1.* Adapted from Figure 2 in "Generalization in Vision and Motor Control" by T. Poggio and E. Bizzi, 2004, *Nature, 431,* p. 769. Copyright 2004 by Macmillan Publishers Ltd. Adapted by permission from Macmillan Publishers Ltd. A schematic diagram of a model of face identification first developed by Riesenhuber and Poggio (1999). Units in the first two layers of the network respond to lines in much the same way as simple and complex cells in V1. Units at higher levels pool information from the previous layer and respond to more complex images in a highly specific manner. For example, the units in layer infratemporal (IT) pool information from posterior IT (PIT) and are tuned to specific faces in specific orientations, and units in AIT are tuned to specific faces across a range of orientations. Although a familiar face can be identified by the activation of a single unit in AIT, unfamiliar faces can only be identified on the basis of a pattern of activation across a set of AIT units.

represented in a localist format at the word level but in a distributed manner at the letter layer. For instance, the word *trap* is thought to be represented locally by a single active unit at the word level but as a pattern of activation over the units *t*, *r*, *a*, and *p* at the letter level. On this definition, even if the word *trap* selectively activates one word unit (with no activation extending to units that code for form similar words), the word *trap* would still be represented in a distributed format at the letter level.

To illustrate their point, Plaut and McClelland (2010) considered the role of the retina in object identification. They claimed that the retina contains all the relevant information for object identification and that objects are represented in a distributed format at this level (e.g., through the coactivation of photoreceptors or perhaps ganglion cells). At the same time, they noted that object knowledge is encoded at a higher level of the visual processing system (e.g., IT), and the information in the retina must be rerepresented at the higher levels before it can access this object knowledge. Quian Quiroga and Kreiman (2010) make a similar point: They state that the retina represents the object implicitly in a distributed format, whereas object knowledge is coded explicitly at higher levels of the visual processing stream.

In the target article, I devoted a large section of the article outlining an argument as to why I think it is a mistake to consider coactive letter representations a distributed representation of a

word (see Bowers, 2009, pp. 223–224). Plaut and McClelland (2010) rejected my analysis on the basis of on the claim that I have failed to distinguish between the knowledge of a word that is coded in the long-term connection weights and an internal representation or an interpretation of a word that is coded as a pattern of activation over units in a network. In addition, they claim that I have adopted an unconventional definition of localist versus distributed representations, focusing on long-term knowledge rather than on the activation of internal representations. In their view, the critical question is whether the units are activated in a localist way or a distributed way.

I agree that it is important to distinguish between the long-term knowledge found in the connection weights of a network and the temporary activation of units in a network. I also agree that previous claims regarding the sparseness of neural representations have focused on the pattern of activity of neurons rather than on the pattern of connection weights. Indeed, this is the perspective I have adopted. That is, I am claiming that familiar words, objects, and faces might plausibly be identified through the activation of single neurons (or collections of redundant neurons). What is unconventional is Plaut and McClelland's (2010) terminological distinction between knowledge and internal representation. According to standard usage, an object can be represented in long-term memory (in connection weights and associated units) or represented in short-term memory (through the temporary activation of its units). On the standard view, an object is identified when the relevant units are sufficiently active.

Regardless of this terminological point, there is a problem with Plaut and McClelland's (2010) analysis. Consider again their example of the retina. Plaut and McClelland claimed that all the information required to identify an object is encoded in the retina but that "the information must be *rerepresented* [italics added] by a hierarchy of visual areas before it can effectively engage object knowledge" (Plaut & McClelland, 2010, p. 285). But a great deal of the information required to identify an object is not in the retina, including the knowledge that allows us to derive a three-dimensional percept from a two-dimensional image, knowledge that supports perceptual grouping, knowledge that derives shape from shading, and so on. The higher levels of the system do not simply rerepresent information projected on the retina but also add new information. In the same way, new information is encoded at the word level of the IA model, namely, the knowledge that the letters *t*, *r*, *a*, and *p* go together to form a familiar word. This information is not at the level of the letter units, and a pattern of activation across these units does not capture (or represent) this information. It does no good to say that object knowledge is encoded implicitly in the retina, as again, key information about objects (and words) is not encoded in the retina at all.

To summarize, the term *distributed representation* can be defined so broadly that every model includes distributed representations, including the archetypal localist model in psychology, namely, the IA model. But doing this renders the term meaningless: Everyone agrees that many neurons (units) are coactivated in each cognitive act. If Poggio and Bizzi (2004) want to describe their (impressive) model as a distributed model even though it implements an explicit extension of the Hubel and Wiesel (1968) hierarchy and works very much like the IA model of word identification, then they are only rejecting a straw-man version of the grandmother cell theory.

Although these definitional and conceptual issues are important, there is one sense in which they do not matter. PDP modelers have rejected the IA model because it includes the wrong types of representations. Accordingly, even if I were to concede that the IA model and the Poggio and Bizzi (2004) model include distributed representations of familiar words and faces (which I do not), the representations in these models lend no support to the PDP approach.

## Distributed Representations and PDP Models

What sorts of distributed representations are PDP modelers committed to? In the target article, I note that PDP modelers have relied on two different sorts, namely, sparse representations that support rapid learning (but cannot generalize) and dense distributed representations that support generalization (but cannot learn quickly). Plaut and McClelland (2010) make the important point that the space of representations can vary along two dimensions: *sparsity*, defined as the proportion of units activated in response to a single object, and *perplexity*, defined as the degree to which a unit responds to dissimilar objects. In principle, units in a network can take on all levels of sparsity and perplexity. For example, an object could activate a single unit in a network (the most extreme version of sparseness), and at the same time, this unit might respond to a number of different and unrelated inputs (that is, the response of this one unit is highly perplexing). Alternatively, an object could coactivate many units (a nonsparse code), and at the same time, each unit could uniquely respond to this object (the units are maximally interpretable). Grandmother cell models tend to include both sparse and interpretable representations, but the combination of nonsparse and interpretable representations also constitute a grandmother cell theory (a grandmother cell theory with redundancy). When defining grandmother cells, it is the interpretability of units that matters. For similar analysis (and conclusions) see Földiák (2009).

According to Plaut and McClelland (2010), I have an overly constrained view of the types of representations that PDP models can learn, and under the relevant conditions, these models can learn all variety of internal representations within this two dimensional space,[1] including units whose responses are highly sparse and interpretable. Furthermore, Plaut and McClelland (2010) claimed that the interpretability of internal representations is irrelevant to their theoretical approach. That is, there is no commitment to dense distributed representations. On their view, the many reports that neurons respond in a highly selective manner pose no challenge to the PDP approach.

But the claim that the PDP approach is not committed to dense distributed representations seems at odds with many past statements of PDP theorists. For example, it is inconsistent with the claims of Elman (1995) and Smolensky (1988), who wrote that hidden units cannot be interpreted one at a time (see Bowers, 2009, p. 226). In the same way, Rogers and McClelland (2004) seem to

[1] Although PDP models may learn internal representations that capture all variety of sparseness and perplexity, PDP models do not learn context independent representations that play a role in symbolic systems. This leads to limitations in their ability to generalize (cf., Bowers, Damian, & Davis, 2009; Bowers & Davis, 2009; but see Botvinick & Plaut, 2009; Sibley, Kello, Plaut, & Elman, 2009).

endorse the view that single hidden units in PDP networks are uninterpretable. When describing a PDP model of semantic memory originally developed by Rumelhart and contrasting this model with alternative (non-PDP) connectionist approaches, they wrote,

> An important departure from other representational schemes (including those used in some other connectionist approaches) is that the internal representations acquired by the Rumelhart network are not lists of semantic features—in no way are they directly interpretable semantically. Individual units in the model's Representation layer do not encode the presence or absence of explicit, intuitive object properties. . . . The individual semantic properties can only be recovered through the combined effects of the units in the distributed representation, working in concert with units in other parts of the network. (Rogers and McClelland, 2004, pp. 62–63)

That is, not only are hidden units in PDP models uninterpretable, but this constitutes an important theoretical contrast with alternative neural network approaches. I would suggest that most PDP theorists adopt this perspective, which helps to explain why there have been so few attempts to study the responses of hidden units one at a time.

### Are Grandmother Cell Theories Really Plausible?

Of course, even if the neuroscience rules out the dense distributed representations associated with the PDP framework, it does not follow that localist models are right, or even plausible. Plaut and McClelland (2010) claimed that there are fundamental conceptual and computational problems with localist theories, and Quian Quiroga and Kreiman (2010) argued that various findings in neuroscience rule out grandmother theories. If this is the case, some other form of distributed coding would be implicated, such as coarse or sparse coding. But there are reasons to challenge these analyses.

### A Computational Challenge to Grandmother Cell Theories

The core claim of a grandmother theory is that single neurons at the top of a hierarchy represent one familiar thing, be it an object, face, or word. Plaut and McClelland (2010) argued that this approach is subject to a computational limitation. That is, although this form of coding might (in principle) support object identification at a basic level (e.g., a generic cat, rat, or car), the task of perceiving the intricate details of scenes, including the specific tulip on McClelland's dining room table, is thought to be untenable. The problem is that it results in a combinatorial explosion: There is no way that unique grandmother cells can be devoted to each possible discriminable state in the world (e.g., a specific tulip in a particular context, a specific piece of toast complemented with Marmite, etc.). Plaut and McClelland also considered instance theories of perception, in which each token of an object is stored with its own localist representation. Here, the brain does not need to include separate grandmother cell for an infinity of possible discriminable states but need include only the states that have been experienced directly. Although this might at first appear to provide a computationally tractable version of a grandmother cell theory, Plaut and McClelland claimed that instance theories are incompatible with the localist theories. As they noted, on an instance theory, the identification of a piece of toast involves the coactivation of many toast nodes.

The core problem with this analysis is that a grandmother cell theory is only committed to the claim that single neurons code for an equivalent class of familiar things. Accordingly, it is only necessary to devote a single unit to a specific tulip on McClelland's dining room table if McClelland can identify it (as opposed to other tulips). Barring this, it is possible that there is a unit for tulips (in general) at the top of his visual processing hierarchy, and a tulip is identified when the tulip cell is activated beyond some threshold. The perceptual vividness of each tulip might be due to the specific set of coactive neurons across all the levels of the visual hierarchy. This is only a concession to the distributed approach if distributed coding is defined as multiple neurons firing.

This characterization of vision is consistent with the reverse hierarchy theory of vision, in which the conscious perception of a visual scene first occurs at the top of the visual hierarchy. The representations at this level support the perception of the "gist of a scene," but not its fine details. The representations at lower levels of the hierarchy support the conscious perception of these details, but this requires the observer to scrutinize a scene (cf. Ahissar & Hochstein, 2004). This theory of vision helps make sense of the finding that one can identify objects at an abstract level without first perceiving its details (although still one can detect that something is in the visual scene before one knows what it is; cf. Bowers & Jones, 2008). A similar approach was assumed by Poggio and colleagues (e.g., Serre, Oliva, & Poggio, 2007) and others (e.g., Grossberg, 2003), who assumed that the initial categorization of objects is based on abstract representations at the top of a hierarchy, with the vividness of perception relying on the later involvement of lower level representations (perhaps guided by top-down feedback or even a resonance of activation between levels). For additional discussion of this issue, see the target article (Bowers, 2009, pp. 223–224).

In one respect, grandmother cells at the top of a visual hierarchy are quite different from the representation within an instance theory. That is, these grandmother cells code for information in an abstract format. Nevertheless, the representations within an instance theory might also be considered grandmother cells, given that individual units are interpretable (e.g., a single unit codes for a specific piece of toast). In fact, it could be argued that the key difference between theories is that there is an extra visual level in classic grandmother theories in which abstract units in layer $N$ pool information from specific units in layer $N - 1$. For example, consider Figure 1, in which single units in AIT code for specific faces independent of orientation (grandmother cells), and single units in IT code for specific faces in specific orientations (what might be called instance units). An instance version of this model might attempt to due away with the final (abstract) level of the visual hierarchy, and mediate abstraction through the coactivation of familiar instance units. But the critical point for present purposes is that the top units in this instance version of the theory are also interpretable. The fact that multiple instance units are activated in response to an input does not rule it out as a grandmother cell theory (just as this observation does not rule out a classical grandmother cell theory that includes abstract localist units on top).

## Empirical Challenges to Grandmother Cell Theories

In the target article (Bowers, 2009), I reviewed a number of studies that identified neurons in IT cortex and in the medial temporal lobe (MTL) that responded to images in a highly selective fashion, sometimes responding to one image out of many. In one study, the response of single IT neurons in monkeys could account for human performance in a face perception and/or memory task (Keysers, Xiao, Foldiak, & Perrett, 2001). I concluded that these findings falsify existing PDP models that learn dense distributed representations and that future research is required to distinguish between grandmother and highly sparse coding schemes in which each unit is involved in coding more than one thing.

However, Quian Quiroga and Kreiman (2010) argued that the grandmother cell hypothesis is falsified based existing data. They summarized the conclusions of Waydo, Kraskov, Quian Quiroga, Fried, and Koch (2006), who carried out a Bayesian analysis on the responses of 1,425 MTL neurons in humans (based on data collected by Quian Quiroga et al., 2005). Many of these neurons did not respond to any images, other neurons responded to one or very few images, and still others responded to many images (e.g., 25% of the images displayed). Despite the selectivity of some neurons, they concluded that each image activated approximately 5 million neurons within MTL (a measure of population sparseness), and each neuron is involved in representing between 50 and 150 objects (a measure of lifetime sparseness). On the basis of both sets of estimates, Waydo et al. (2006) rejected grandmother cells.

It is important to describe how these numbers were reached. The Bayesian analysis relied on the response of all 1,425 neurons, and it provided an estimate of the number of images that the typical neuron in MTL responds to (in the form of a probability distribution across a range of possible values). On the basis of this analysis, Waydo et al. (2006) concluded that each neuron responds to approximately 0.54% of the distinct objects in the world (when responding was defined as 5 *SD*s above baseline). Given that there are approximately $10^9$ neurons in MTL, they concluded that $0.0054 \times 10^9$ or ~5.4 million neurons responded to the image. And on the assumption that people know between 10,000 and 30,000 objects, each individual neuron will be involved in representing $0.0054 \times 10{,}000$ to $0.0054 \times 10{,}000$ (or 50 to 150) images.

There are both analytical and conceptual problems with these conclusions. The computational analyses rest on the key assumptions that all the neurons in MTL are involved in identifying basic level things (from a universe of between 10,000 and 30,000 things), that all neurons have the same lifetime sparseness value, and that each thing is equally well-represented in MTL. But all these premises are suspect. For example, many neurons may be involved in coding for the familiarity (as opposed to the identity) of an object. Such neurons would fire to a wider range of images than would neurons involved in coding the identity of a specific person or object. In other words, the lifetime sparseness of neurons in the MTL might not be the same, and it would be a mistake to average the sparseness of all the neurons in the study to reach an estimate of the sparseness of neurons coding for specific persons or objects. In addition, there is no reason to assume that all objects are equally well represented. Quian Quiroga et al. (2005) interviewed the patients before the recordings and selected images of persons and things that were highly familiar. According to the multiple trace theory, important memories in the hippocampus are redundancy coded (massively so), and this provides a parsimonious account of the retrograde amnesia that is observed following damage to the hippocampus (Nadel & Moscovitch, 1997). Looking for neurons that respond to highly familiar objects increases the likelihood that the object is redundantly coded, and it is presumably easier to find a neuron that responds to an object if it is from a redundant set of neurons that all respond to the same object. That is, looking for neurons that respond to highly familiar objects may result in reduced estimates of sparseness.[2] In addition, on a technical level, it is more difficult to identify neurons with higher levels of sparseness than to identify neurons with lower levels of sparseness (cf. Shoham, O'Connor, & Segev, 2006). So again, the analysis is biased against grandmother cells. Even the Bayesian estimate of the range of possible sparseness values is biased against grandmother cells. Lifetime sparseness is defined as the percentage of images that a neuron responds to, and a grandmother sparseness value is close to 0 (e.g., 1/30,000 images). In their Bayesian analysis, Waydo et al's (2006) a priori estimate of sparseness (before any data were collected) was that all values of sparseness from 0 to 1 are equally likely. That is, on their analysis, it is a priori highly unlikely that a neuron is a grandmother cell. The consequence is that their estimates of the range of sparseness values around the mean are biased against grandmother cells. Some of these points are acknowledged by Quian Quiroga and Kreiman (2010; also by Waydo et al., 2006), but it is important to highlight that these assumptions could easily change the estimates of population and lifetime sparseness by an order of magnitude (or more).

Equally problematic, Waydo et al. (2006) defined grandmother cells both on the basis of lifetime sparseness (a grandmother cell should respond to only one object) and on population sparseness (a single neuron should fire in response to an image). However, as noted above, the only relevant criterion is lifetime sparseness (or perplexity)—a grandmother cell theory is still supported if many neurons can be interpreted unambiguously in the same (redundant) way. By including population sparseness in their definition of grandmother cells, Waydo et al. (2006) have adopted the wrong criterion that makes rejection of grandmother cells guaranteed. Imagine that Quian Quiroga et al. (2005) had recorded from a million neurons and estimated that neurons respond to 1/30,000 images (for which 30,000 is the estimate of the number of known objects). On a lifetime measure of sparseness, this would constitute a grandmother cell. Nevertheless, Waydo et al. (2006) would conclude that $1/30{,}000 \times 10^9$ or ~30,000 neurons respond to a given image. This is far less than their estimate that 5 million neurons respond to a given object (which they take to be inconsistent with grandmother cells), but 30,000 coactive neurons is also inconsistent with the theory, if it is assumed that one and only one neuron should respond to a given object.

[2] Although measuring neural responses to highly familiar images may introduce a bias against high estimates of sparseness, my previous claim that Waydo et al.'s (2006) calculations are equally consistent with the conclusion that there are 50–150 redundant neurons for each object or person is incorrect, as noted in both commentaries.

Despite these problems, the above analyses might still appear to rule out grandmother cells. That is, even if Waydo et al's (2006) estimate of lifetime sparseness is off by an order of magnitude for the reasons described above, the conclusion that each neuron is involved in representing between 5 and 15 things (as opposed to 50–150 things) would falsify grandmother theories.

However, the fact that a neuron fires in response to an image of an object does not imply that the neuron is involved in representing the object. As detailed in the target article (Bowers, 2009), localist units are often activated by things they do not represent. That is, units are incidentally activated, by virtue of similarity. For example, as noted in Figure 4 of the target article, the *blue* unit in the IA model is activated in response to the word *blur*, but this unit is not involved in representing *blur*. In the same way, the finding that a given neuron responds to more than one thing does not imply that it is involved in representing multiple things. The observation that a neuron responds most strongly to one thing and more weakly to similar things is consistent with a grandmother cell theory.

## Is This Account of a Grandmother Cell Theory Unfalsifiable?

I am arguing that the visual system might work something like the IA model of word identification or the Poggio and Bizzi (2004) model of face identification, in which familiar stimuli (e.g., words or faces) can be identified on the basis of a single unit passing some threshold of activation. By my definition, these are grandmother cell theories, even though a given unit responds to more than one familiar word or face (by virtue of similarity), and I am allowing for the possibility that there may be substantial redundancy, with many units representing the same thing. Does this make my account of grandmother cell theories unfalsifiable?

Quian Quiroga and Kreiman (2010) argued that it does. They pointed out that researchers do not know what a neuron "codes for" as opposed to what it "responds to" and concluded that it is implausible to reject a distributed coding scheme when there is direct evidence that single neurons respond to multiple things. Indeed, they take this evidence to be inconsistent with grandmother cells. The problem with this conclusion is that single units in localist networks also respond to multiple things. I agree that this makes it difficult to distinguish between these theoretical approaches, but the important differences between these positions remain. The proper response is not to reject grandmother cells (or distributed coding) but rather to delay any strong conclusions until more telling data are obtained.

One obvious way forward is to test highly selective neurons with many more images and find out whether indeed they respond to other images (at a similar level) or whether they have a clear preference for one stimulus among thousands. Ideally, the studies should be carried out in IT rather than MTL because the grandmother cell theory is concerned with how the perceptual system works, not episodic memory. Additional approaches to tackling this question are possible as well. For example, Nicholas and Newsome (2002) took a different approach to address this question in the domain of motion perception (with mixed results; see Figure 12 from target article, Bowers, 2009). And of course, the relative successes of localist and distributed models in accounting for behavior will speak to this issue as well. Future work may indeed rule out grandmother cells, but at present, the strong conclusion of Quian Quiroga and Kreiman (2010) is premature.

## Conclusions

There is an extensive literature of single-cell recording studies, and the uncontroversial conclusion is that the visual system is organized into a hierarchy of processing steps and that neurons at the top of the hierarchy (in IT) often respond to stimuli in a highly selective fashion. The first main conclusion I draw is that this literature is inconsistent with many existing PDP models. PDP models that learn via back propagation do not learn this hierarchical structure, and the response of single units is generally highly nonselective. Most importantly, these data are inconsistent with the more general theory of cognitive processing that most PDP modelers have adopted, namely, that words, objects, and faces are identified on the basis of dense distributed representations in the cortex.

My second main conclusion is that the widespread rejection of grandmother cell theories in neuroscience is unjustified, or at least premature. Neuroscientists have properly rejected theories in which it is assumed that a single neuron codes for a complex scene (e.g., a weeping grandmother) and theories that assume that one and only one neuron fires in response to an object or a face. But this is an overly constrained view of a grandmother theory. My premise is that grandmother cell theories in neuroscience are the equivalent of localist models in psychology, and localist models make neither of these claims. Indeed, computational neuroscientist have developed models of face identification based on single cell recording data that are very much like the IA model of word identification (e.g., Riesenhuber & Poggio, 1999). In fact, these models provide an explicit extension of the hierarchical model of vision first proposed by Hubel and Wiesel (1968), with localist representations of words and faces on top. This type of grandmother theory is plausible given the neuroscience to date.

## References

Ahissar, M., & Hochstein, S. (2004). The reverse hierarchy theory of visual perceptual learning. *Trends in Cognitive Sciences, 10,* 457–464.

Botvinick, M., & Plaut, D. C. (2009). Empirical and computational support for context-dependent representations of serial order: Reply to Bowers, Damian, and Davis (2009). *Psychological Review, 116,* 998–1002.

Bowers, J. S. (2009). On the biological plausibility of grandmother cells: Implications for neural network theories in psychology and neuroscience. *Psychological Review, 116,* 220–251.

Bowers, J. S., Damian, M. F., & Davis, C. J. (2009). A fundamental limitation of the conjunctive codes learned in PDP models of cognition: Comments on Botvinick and Plaut (2006). *Psychological Review, 116,* 986–997.

Bowers, J. S., & Davis, C. J. (2009). Learning representations of word-forms with recurrent networks: Comment on Sibley, Kello, Plaut, & Elman (2008). *Cognitive Science, 33,* 1183–1186.

Bowers, J. S., & Jones, K. W. (2008). Detecting objects is easier than categorizing them. *Quarterly Journal of Experimental Psychology, 61,* 552–557.

Elman, J. L. (1995). Language as a dynamical system. In R. F. Port & T. van Gelder (Eds.), *Mind as motion: Explorations in the dynamics of cognition* (195–223). Cambridge, MA: MIT Press.

Földiák, P. (2009). Neural voding: Nonlocal but explicit and conceptual. *Current Biology, 19,* R904–R906.

Grossberg, S. (2003). Bring ART into the ACT. *Behavioral and Brain Sciences, 26,* 610–611.
Hubel, D. (1995). *Eye, brain, and vision.* New York, NY: Scientific American Library.
Hubel, D. H., & Wiesel, T. N. (1968). Receptive fields and functional architecture of monkey striate cortex. *Journal of Physiology: London, 195,* 215–243.
Keysers, C., Xiao, D. K., Foldiak, P., & Perrett, D. I. (2001). The speed of sight. *Journal of Cognitive Neuroscience, 13,* 90–101.
McClelland, J. L., & Rumelhart, D. E. (1981). An interactive activation model of context effects in letter perception. 1. An Account of Basic Findings. *Psychological Review, 88,* 375–407.
Nadel, L., & Moscovitch, M. (1997). Memory consolidation, retrograde amnesia and the hippocampal complex. *Current Opinion Neurobiology, 7,* 217–227.
Nicholas, M. J., & Newsome, W. T. (2002). Middle temporal visual area microstimulation influences veridical judgments of motion direction. *Journal of Neuroscience, 22,* 9530–9540.
Page, M. P. A. (2000). Connectionist modeling in psychology: A localist manifesto. *Behavioral and Brain Sciences, 23,* 443–512.
Plaut, D. C., & McClelland, J. L. (2010). Locating object knowledge in the brain: Comment on Bowers's (2009) attempt to revive the grandmother cell hypothesis. *Psychological Review, 117,* 284–290.
Poggio, T., & Bizzi, E. (2004, October). Generalization in vision and motor control. *Nature, 431,* 768–774.
Quian Quiroga, R., & Kreiman, G. (2010). Measuring sparseness in the brain: Comment on Bowers (2009). *Psychological Review, 117,* 291–299.
Quian Quiroga, R. Q., Reddy, L., Kreiman, G., Koch, C., & Fried, I. (2005, June). Invariant visual representation by single neurons in the human brain. *Nature, 435,* 1102–1107.
Riesenhuber, M., & Poggio, T. (1999). Hierarchical models of object recognition in cortex. *Nature Neuroscience, 2,* 1019–1025.
Rogers, T. T., & McClelland, J. L. (2004). *Semantic cognition: A parallel distributed processing approach.* Cambridge, MA: MIT Press.
Seidenberg, M. S., & Plaut, D. C. (2006). Progress in understanding word reading: Data fitting versus theory building. In S. Andrews (Ed.), *From inkmarks to ideas: Current issues in lexical processing.* Hove, England: Psychology Press.
Serre, T., Oliva, A., & Poggio, T. (2007). A feedforward architecture accounts for rapid categorization. *Proceedings of the National Academy of Sciences, USA, 104,* 6424–6429.
Shoham, S., O'Connor, D. H., & Segev, R. (2006). How silent is the brain: Is there a "dark matter" problem in neuroscience? *Journal of Comparative Physiology: A. Neuroethology Sensory Neural and Behavioral Physiology, 192,* 777–784.
Sibley, D. E., Kello, C. T., Plaut, D. C., & Elman, J. L. (2009). Sequence encoders enable large-scale lexical modeling: Reply to Bowers and Davis (2009). *Cognitive Science, 33,* 1187–1191.
Smolensky, P. (1988). Putting together connectionism—Again. *Behavioral and Brain Sciences, 11,* 59–70.
Waydo, S., Kraskov, A., Quian Quiroga, R. Q., Fried, I., & Koch, C. (2006). Sparse representation in the human medial temporal lobe. *Journal of Neuroscience, 26,* 10232–10234.

Received July 6, 2009
Revision received September 21, 2009
Accepted September 22, 2009 ■

## Postscript: Some Final Thoughts on Grandmother Cells, Distributed Representations, and PDP Models of Cognition

Jeffrey Bowers
*University of Bristol*

Below, I briefly respond to a number of terminological, theoretical, and empirical issues raised in some postscripts. The goal is not to respond to each outstanding point but rather to address some comments that in my view confuse rather than clarify matters. I respond to Plaut and McClelland (2010) and Quian Quiroga and Kreiman (2010) in turn.

According to Plaut and McClelland (2010), the parallel distributed processing (PDP) approach is defined by its commitment to interactivity and graded constraint satisfaction. Many localist models, including the interactive activation (IA) model, are characterized in this way, and accordingly, they write that "it makes perfect sense to speak of localist PDP models" (p. 289). On this definition, any evidence in support of grandmother cells constitutes a challenge not to the PDP approach per se, just to models that include distributed representations. This characterization of the PDP approach constitutes more of a terminological point than a theoretical point, but it is worth noting that it is inconsistent with many previous statements in which distributed representations are described as a core principle (e.g., Plaut & Shallice, 1993; Seidenberg, 1993). Furthermore, this definition renders the PDP approach so broad that it encompasses almost all neural networks, including network models that are typically seen as inconsistent with the PDP framework (e.g., Grossberg, 1980; Davis, 1999; Hummel & Biederman, 1992). If advocates of the PDP approach are only committed to interactivity and graded constraint satisfaction, with no commitment to the form of the representations that underpin cognition, then there is nothing unique (or novel) about the approach per se.

Even in the context of this broad definition, Plaut and McClelland (2010) argued that my version of a localist model is inconsistent with the PDP approach. That is, I am advocating models in which word, object, and face identification is achieved when a localist representation is activated beyond some threshold. This is said to undermine the key successes of localist PDP models which rely on cascaded processing. For instance, they note that the IA model can explain context effects in letter perception (e.g., a facilitation in identifying a letter embedded in a pseudoword) with the assumption that partial and ambiguous activity at the letter level propagates forward to the word level and partial and ambiguous activity at the word level feeds back to the letter level (although feedback is not strictly necessary to account for the context effects; Grainger & Jacobs, 1996). These context effects in the IA model are observed without thresholds (or identifying any words), and indeed, according to Plaut and McClelland (2010), the inclusion of thresholds would undermine a model's ability to account for the effects.

Plaut and McClelland (2010) appear to have mistaken my comments regarding thresholds with the claim that processing is discrete; that is, when partial activation of letters and words cannot be passed on to subsequent levels and can play no role in processing. In both the target article (Bowers, 2009) and my reply (Bowers, 2010), I

describe localist models in which a given input coactivates multiple units and in which the competition between coactive units plays a role in selecting the target. That is, the competition serves to restrict the number of units that pass some threshold. Thresholds and cascadedness are orthogonal issues, and accordingly, a model with thresholds can account for letter context effects in word perception. Indeed, as noted by Plaut and McClelland (2010), thresholds are often implemented in the IA model. The important point for present purposes is that thresholds in a network in no way undermine the distinction between a unit that codes for an input (e.g., a unit that codes for the word *blue*) and a unit that is only incidentally activated by virtue of form similarity (e.g., a unit coding for *blur* responding to the input *blue*). Equally important, this is all tangential to the question of whether a localist model (PDP or otherwise) is biologically plausible.

Plaut and McClelland (2010) also raised the concern that localist models have no ready way to assign units to inputs. How is the model to know whether a unit should be assigned to a particular grandmother as opposed to grandmothers in general? Or tulips in general as opposed to a particular tulip? What constitutes an equivalence class? They claimed that there are no well-developed learning theories to address these difficult problems and suggested that they may well be intractable for localist approaches in principle. But there are existing implemented localist models that show some promise in addressing these issues. For example, adaptive resonance theory (ART) models of Grossberg (1980, 1987) can learn localist representations at various levels of abstraction. A critical property of these networks is that they include a vigilance parameter that directly affects the granularity of the learned categories. The vigilance parameter is adjusted based on the feedback. If a model makes a mistake in categorizing an input (e.g., categorizing a random old lady as my grandmother or an early blooming tulip as a late bloomer), the vigilance is set higher, and as a consequence, the model learns to categorize perceptually similar inputs with separate localist units. The vigilance parameter also plays a key role in addressing the stability–plasticity dilemma, such that learning new categories (e.g., learning that this specific face belongs to my grandmother) does not erase old knowledge (that my grandmother is an old woman). As a consequence, the model does not have to decide whether to code information at either an abstract or a specific level—it can do both.

Other localist models might be developed to address these concerns as well. For example, consider the model of face identification developed by Riesenhuber and Poggio (1999). A key feature of this model is its hierarchical structure, in which information is coded at various levels of abstraction. For instance, in one layer of the network, the model includes localist units that code for specific views of familiar persons, and in a subsequent layer, units code for familiar persons independent of viewpoint. So once again, the model does not have to choose whether to code a familiar object at an abstract or specific level because it can do both. The Riesenhuber and Poggio (1999) model does not learn, but it is not implausible to imagine a learning algorithm that develops more levels of localist coding as a function of expertise. Just as we are all experts in face recognition and can distinguish one grandmother from another, a florist can distinguish different types of tulips. In both cases, this might be accomplished by the recruitment of localist representations at a subordinate level (in addition to separate units at a basic level). Of course, neither of these models provides a complete answer to these challenging questions (nor do distributed PDP models), but claims regarding the computational limitations of localist models seem premature.

With regards to the neuroscience, Quian Quiroga and Kreiman (2010) highlight that most neurons in their studies responded to more than one image. Even some of the most selective neurons with the medial temporal lobe (MTL) responded to more than one thing—for example, a neuron that fired to two basketball plays, another to two different landmarks, and yet another that responded to Luke Skywalker and Yoda (characters in Star Wars), among other examples. Nevertheless, a few neurons responded robustly to only one out of all the images tested, and the catalogue of examples is expanding. For example, Quian Quiroga, Kraskov, Koch, and Fried (2009) reported a single neuron in MTL that responded to a written word, spoken word, or image of Saddam Hussein but responded to no other stimulus in the experiment. What is to be made of the mixed set of results? Does the fact that most of these neurons responded to more than one image compromise the grandmother cell hypothesis? More generally, is the grandmother cell hypothesis falsified by the Bayesian analysis reported by Waydo, Kraskov, Quian Quiroga, Fried, and Koch (2006) that demonstrates that a given image will inevitably activate many neurons in MTL and that each of these neurons will inevitably respond to many images? I would suggest not. The critical point that needs to be reemphasized is that the units in localist models respond in a similar way; namely, each localist unit responds to more than one input, and a given input activates more than one unit. That is, lifetime sparseness and population sparseness in both localist models and the MTL are extremely high but are still not at the limit of sparseness. The analysis of Waydo et al. (2006) is an important way forward in characterizing the response profiles of neurons in the MTL, but given the range of possible estimates of these measures at present, it is not appropriate to reject localist representations (or grandmother cells) on the basis of their data just yet. What would falsify a grandmother cell theory is an estimate of lifetime and population sparseness in IT that falls outside the range of plausible values for localist models.

This relates to a more fundamental problem with Quian Quiroga and Kreiman's (2010) position. When they rejected the distinction between what a neuron "codes for" and what it "responds to," they are rejecting a fundamental distinction between localist and distributed networks. By ignoring this distinction, they only end up rejecting a straw-man version of a grandmother cell theory. Our impasse on this point might reflect a confusion of terminology between disciplines, and it might be helpful to put the issue in another way. Consider again the neurobiological model of face perception by Riesenhuber and Poggio (1999), inspired by single cell recording data. In this model, individual units are tuned to respond to specific familiar faces, and at the same time, a specific input activates more than one face unit (the target face and units tuned to other similar faces). On my definition, this constitutes a localist model in which each unit represents one specific face (and does not contribute to the representation of other faces). To see this, consider what would happen to the identification of a familiar face if all the coactive units were removed from the network (apart from the unit tuned to the target). The answer is that the model would continue to recognize the face just fine. Conversely, if this one unit was removed from the network, the model would fail to recognize the input as a familiar face. This raises the following

question: Do Quian Quiroga and Kreiman (2010) take their data as inconsistent with this modeling approach? If not, we are essentially in agreement—single cell recording data are consistent with models that work very much like the localist models in psychology.

Finally, Quian Quiroga and Kreiman (2010) reiterated their claim that all the information required for object recognition is in the retina, but in a distributed and implicit code. They reject my claim that a great deal of information required to identify words, objects, and people is located outside the retina, in higher levels of the visual processing pathway. This is said to violate a data processing inequality, according to which processing cannot add information. But there is something wrong with this characterization of the processing inequality. It is clear is that the retina does not include the information about what letter strings constitute words or what configuration of active ganglion cells constitutes an image from my grandmother. This information is stored in higher levels of the visual system, acquired through experience. Similarly, evolution may have endowed higher level visual systems with computational principles to derive shape from shading, depth cues, and so on. Although I agree that processing (transforming) information in and of itself cannot add new information (all transformations are by definition derivable from the input), bottom-up information can nevertheless access other databases of knowledge that contain new information that cannot be derived from the input alone. For example, on viewing a duck, I can predict that the duck might quack. This surely does constitute a violation of data processing inequality.

To conclude, I make one observation that should prove uncontroversial. Regardless of one's position regarding the localist–distributed debate, the target article (Bowers, 2009) highlights a promising approach for evaluating network models in the future, namely, exploring the responses of hidden units one at a time in response to a wide range of inputs. There is a striking disconnect between the methods of neurophysiology, in which neurons are studied one at a time, and the methods in cognitive science, in which hidden units in PDP models are generally studied in combination. This disconnect constitutes a missed opportunity to provide some important constraints on theorizing. An analysis of single units may provide some insights into the conditions under which different coding schemes emerge in neural network models and some insights into why the brain adopts the solutions it does. These analyses might even show that localist representations are required to solve some fundamental computational tasks in perception and memory.

## References


Bowers, J. S. (2009). On the biological plausibility of grandmother cells: Implications for neural network theories in psychology and neuroscience. *Psychological Review, 116,* 220–251.

Bowers, J. S. (2010). More on grandmother cells and the biological implausibility of PDP models of cognition: A reply to Plaut and McClelland (2010) and Quian Quiroga and Kreiman (2010). *Psychological Review, 117,* 300–308.

Davis, C. J. (1999). The self-organising lexical acquisition and recognition (SOLAR) model of visual word recognition (Doctoral dissertation, University of New South Wales, Sydney, New South Wales, Australia, 1999). *Dissertation Abstracts International, 62,* 594.

Grainger, J., & Jacobs, A. M. (1996). Orthographic processing in visual word recognition: A multiple read-out model. *Psychological Review, 103,* 518–565.

Grossberg, S. (1980). How does a brain build a cognitive code? *Psychological Review, 87,* 1–51.

Grossberg, S. (1987). Competitive learning: From interactive activation to adaptive resonance. *Cognitive Science, 11,* 23–63.

Hummel, J. E., & Biederman, I. (1992). Dynamic binding in a neural network for shape recognition. *Psychological Review, 99,* 480–517.

Plaut, D. C., & McClelland, J. L. (2010). Postscript: Parallel distributed processing in localist models without thresholds. *Psychological Review, 117,* 284–290.

Plaut, D. C., & Shallice, T. (1993). Deep dyslexia: A case-study of connectionist neuropsychology. *Cognitive Neuropsychology, 10,* 377–500.

Quian Quiroga, R., Kraskov, A., Koch, C., & Fried, I. (2009). Explicit encoding of multimodal percepts by single neurons in the human brain. *Current Biology, 19,* 1308–1313.

Quian Quiroga, R., & Kreiman, G. (2010). Measuring sparseness in the brain: Comment on Bowers (2009). *Psychological Review, 117,* 291–299.

Riesenhuber, M., & Poggio, T. (1999). Hierarchical models of object recognition in cortex. *Nature Neuroscience, 2,* 1019–1025.

Seidenberg, M. S. (1993). Connectionist models and cognitive theory. *Psychological Science, 4,* 228–235.

Waydo, S., Kraskov, A., Quian Quiroga, R., Fried, I., & Koch, C. (2006). Sparse representation in the human medial temporal lobe. *Journal of Neuroscience, 26,* 10232–10234.