City Research Online

City, University of London Institutional Repository

Citation: Wills, A. J. & Pothos, E. M. (2012). On the adequacy of current empirical evaluations of formal models of categorization. Psychological bulletin, 138(1), pp. 102-125. doi: 10.1037/a0025715

This is the unspecified version of the paper.

This version of the publication may differ from the final published version.

Permanent repository link: https://openaccess.city.ac.uk/id/eprint/1978/

Link to published version: https://doi.org/10.1037/a0025715

Copyright: City Research Online aims to make research outputs of City, University of London available to a wider audience. Copyright and Moral Rights remain with the author(s) and/or copyright holders. URLs from City Research Online may be freely distributed and linked to.

Reuse: Copies of full items can be used for personal research or study, educational, or not-for-profit purposes without prior permission or charge. Provided that the authors, title and full bibliographic details are credited, a hyperlink and/or URL is given for the original metadata page and the content is not changed in any way.

City Research Online: http://openaccess.city.ac.uk/ publications@city.ac.uk

On the adequacy of current empirical evaluations of formal models of categorization

MS #2010-0444

Revised manuscript for Psychological Bulletin

17-Dec-12

Abstract

Categorization is one of the fundamental building blocks of cognition, and the study of categorization is notable for the extent to which formal modeling has been a central and influential component of research. However, the field has seen a proliferation of divergent, non-complementary models with little consensus on the relative adequacy of these accounts. Progress on assessing relative adequacy of formal categorization models against these criteria has, to date, been limited because (a) formal model comparisons are narrow in the number of models and phenomena considered, and (b) models do not often clearly define their explanatory scope. Progress is further hampered by the practice of fitting models with arbitrarily variable parameters to each data set independently. Reviewing examples of good practice in the literature, we conclude that model comparisons are most fruitful when relative adequacy is assessed by comparing well-defined models on the basis of the number and proportion of irreversible, ordinal, penetrable successes (principles of minimal flexibility, breadth, good-enough precision, maximal simplicity, and psychological focus).

Keywords: categorization; cluster; exemplar; model selection; modeling; prototype.

The study of categorization is a fascinating endeavor. The process of constructing and using categories underpins our capacity to encode and apply information in the world in an efficient and competent manner but also, ultimately, our ability to think in terms of abstract ideas, such as justice, love, and happiness. Categories facilitate communication, and facilitate inferences about unobserved properties of objects. What are the mechanisms that correspond to psychological categorization processes? This question has been intensely studied for over fifty years (e.g. Bruner, Goodnow & Austin, 1956), and has led to some of the most sophisticated and influential mathematical, computational, and neuroscientific models in psychology. Indeed, categorization research contains one of the most influential formal models in all of psychology - the Generalized Context Model (Nosofsky, 1984)[1]. Yet there is also a profound divergence among modelers at the most fundamental level. How should categorization models be compared? What is the ideal form of a categorization model? What kind of categorization models should we aim to develop? The lack of consensus regarding such key issues has resulted in categorization research being carried out in increasingly independent strands and this has been inhibiting overall progress in the field. Nosofsky, Gluck, Palmeri, McKinley and Glauthier (1994) wrote, "Recent years have seen an avalanche of newly proposed models of category learning and representation. As such models grow increasingly more sophisticated, there is a need to develop increasingly more rigorous testing grounds so that one may choose among them" (p. 352). Almost 20 years later, progress towards this goal remains limited.

In the current article, we first provide a definition of the term *formal model*, consider the principal advantages of formal modeling over other forms of theorizing, briefly summarize some of the leading formal models of categorization, and assess progress to date on the empirical evaluation and comparison of these models. We then set out the approaches

we believe are most likely to lead to progress in the future. We organize our conclusions in terms of a set of criteria for assessing the relative adequacy of models, and a list of dependent and independent variables that any adequate formal model of categorization should be expected to address. Although our focus is on the formal modeling of categorization, the issues we discuss and the proposals we make are not limited to the field of categorization research. As we outline below, formal modeling has a number of potential advantages, and these advantages are quite general. Similarly, the extent to which formal models deliver those advantages depends on the extent to which the problems and pitfalls considered in this article are avoided. Categorization research has been chosen for consideration in this paper because it is one of the parts of psychology in which formal modeling has featured particularly heavily (of course, not the only area – psycholinguistics research is another example).

Definition of a formal model

A formal model is one that unambiguously specifies transformations from one or more independent variables (IVs) to one or more dependent variables (DVs). In the case of formal models of categorization, one independent variable is category structure and one dependent variable is categorization accuracy (for an illustration, see Figure 1).

-- Figure 1 about here --

The phrase "unambiguously specifies" is critical for the definition of a formal model. By unambiguous specification, we mean that the model must express the nature of the transformations such that, for a given set of inputs and model parameter values, the model's output can be determined with some kind of algorithm. A reasonable proxy for algorithmic determinability is whether the process of determining the model's output from its inputs and parameters could be performed by a computer program without further human intervention.

Note that unambiguous specification is not the same as determinism; a model's output might, for example, be the probability of a particular response. The criterion of unambiguous specification largely excludes models expressed purely in verbal terms and typically involves mathematical expression (or expression in terms that can be unambiguously transformed into algorithmic operations, as in high-level computer programming languages).

The case for formal modeling

As Murphy (2011) points out, formal modeling is not without its disadvantages. Compared to informal theories, developing formal models is more time consuming and, perhaps as a result, is arguably more likely to lead to the neglect of empirical phenomena that lie outside the model's scope. So why model? There are at least six advantages to a formal modeling approach – recognition of problem complexity, deeper insight, ambiguity reduction, model comparison, behavior prediction and the prospect of automated cognition. We discuss each of these potential advantages of formal models over more informal theories below:

Recognition of problem complexity

The experience of many modelers is that attempting to transform an informal theory into a formal model often leads to a recognition that the problem under study is substantially more complex than was immediately apparent. In categorization research (and in other areas of psychology) this is partly because informal theories make extensive use of verbal labels that denote intuitively obvious but computationally complex constructs. For example, in unsupervised classification (category formation in the absence of feedback), the category groupings selected most frequently by participants are often those that are most "intuitive" – but formally specifying what makes them so turns out to be quite complex (Pothos & Bailey, 2009).

Deeper insight

In essence, a formal model is a data-reduction technique. Formal models can be thought of as compressing potentially very large data sets down to a small number of values – the parameters of the model. The extent to which these parameters allow a reconstruction of the data through the architecture of the model is the extent to which the compression is appropriate. As long as the model's parameters are *penetrable*, this compression can lead to insights about empirical data that may not be obvious from the raw data set. We discuss this advantage further in the "penetrable models" section of the article.

Ambiguity reduction

Definitionally, a formal model is one that unambiguously specifies IV-DV transformations. As a consequence, the ability of a formal model to encompass a particular set of empirical findings should be unambiguously determinable, given sufficient information about the state of the IVs that form part of the model's input. Of course, to the extent that there is uncertainty about the empirical phenomenon itself (through, for example, measurement error) there may be uncertainty about a model's ability to encompass that phenomenon.

Behavior prediction and automated cognition

Formal models provide the prospect for prediction of behavior - if we can predict the output of cognitive processes from their input, we may also be able to reproduce aspects of cognition in artificial devices. Formal models are able to contribute to behavior prediction and automated cognition, over and above informal ones, because of the ambiguity reduction that they entail.

Facilitation of theory comparison

If more than one formal theory exists and, for each of those theories, their ability to encompass known sets of empirical findings is unambiguously determinable, then it is

possible to compare the relative adequacy of those models. The potential for unambiguous determinability in formal models – and the inherent difficulty of unambiguous determinability in more informal forms of theorizing – constitutes one of the main advantages of formal modeling.

Divergent, non-complementary models

In this section, we briefly summarize some leading formal models of categorization. We do this in order to (a) illustrate the divergent, non-complementary nature of current models, and (b) to provide a context for our proposals concerning the empirical evaluation of models. The models considered in this section are: the Generalized Context Model, the Nosofsky-Smith-Minda prototype model, SUSTAIN, COVIS, KRES, the Simplicity model, and the Rational Model of Categorization. These models were chosen on the basis of either being highly influential, or encapsulating an important aspect of categorization theory, or making some unique or original contribution. Even within those criteria, there were a number of formal models of categorization that were worthy of inclusion, but which we nevertheless excluded in order to keep this section to a manageable length. Before describing specific models, we outline at a broad level the components such models typically have, and how those components relate to each other.

Template for a Formal Model of Categorization

At a general level, the purpose of categorization models is to organize information from our experience in such a way that it allows, amongst other things, predictions about how new stimuli should be classified. One can distinguish between the representations upon which categorization is based and the mechanisms that support the categorization process. However, such a distinction is not always clear-cut, as the nature of the representations affects the processes that are required to operate upon them and, reciprocally, the choice of processes

affects the forms of representation that are plausible. Throughout this paper, the term "formal model" denotes a specific combination of process and representational assumptions.

-- Figure 2 about here --

Figure 2 shows a broad schema for models of categorization; not all models have all components. Categorization is seldom modeled from a retinal starting point – most modelers assume some form of higher-level input representation of the presented stimulus. The attentionally-modulated information from the input representations activates one or more intermediate representations (e.g. prototypes, exemplars). Information from the intermediate representations activates one or more category representations via an evidential mechanism (e.g. similarity computations). Typically, more than one category representation will be activated to some degree. Equally, a mechanism has to be in place that can guide the formation of new categories. Either way, there is a need for a decision mechanism that turns graded information into a categorical response.

Generalized Context Model

The Generalized Context Model (GCM; Nosofsky, 1984) is the most influential formal model of categorization to date[1]. The GCM is a model with exemplar-based intermediate representations (Figure 2) – in other words, it assumes that categories are represented through the storage of specific examples of members of those categories. Formal exemplar models of categorization can be traced back at least as far as Reed's (1972) average-distance model, but GCM is most directly related to the Context Model (Medin & Schaffer, 1978).

As Figure 3A illustrates, GCM represents stimulus input as points in a multidimensional space, with inter-item similarity considered to be a decreasing function of distance in that psychological space. The position of stimuli in that space is sometimes assumed, but more often derived from the multi-dimensional scaling of stimulus

identification confusion probabilities (Nosofsky, 1986), or pairwise similarity ratings (Palmeri & Nosofsky, 2001). Distance in psychological space is typically calculated with a Euclidean metric (Figure 3B) for integral stimuli, but with a *city-block* metric (Figure 3C) for separable stimuli. This follows part of Garner's (1978) operationalization of the integral-separable distinction (integral stimuli are also, for example, those whose dimensions are difficult to selectively attend; hue and saturation, for example).

-- Figures 3 and 4 about here --

Following Shepard (1958), the function relating distance in psychological space to similarity is typically exponential (Figure 3D). This similarity-distance function formalizes the process assumption that categorical decisions are strongly affected by stimuli that are close in psychological space, with differences in distance becoming less important as distance from the presented item increases. Where stimuli are highly confusable, a Gaussian rather than exponential function is sometimes used; this approximates trial-to-trial variability in stimulus representation (Ennis, 1988).

GCM also has an attentional mechanism. In GCM, selective attention is considered to operate on the dimensions of the psychological space (Figure 3A), with attention to dimension X being represented as a factor by which distances on that dimension are multiplied in order to calculate similarity. Conceptually, and as illustrated in Figure 4, this can be considered as the compression and stretching of psychological space. The inclusion of selective attention in GCM was originally motivated through its ability to provide an account for certain relationships between identification difficulty and categorization difficulty (Nosofsky, 1984, 1986). GCM also has a parameter that allows for the overall expansion or contraction of psychological space (Figure 4C). GCM's formalization of selective attention encapsulates the ideas that (a) categorization accuracy can be enhanced by selective attention,

(b) selective attention to stimulus dimensions is a matter of degree, rather than all-or-none, (c) selective attention occurs at the level of stimulus dimensions.

The output of the GCM is a prediction about the probability with which each of the available category responses will be produced, as a function of the presented stimulus, and the nature of the exemplars presented to the participants. GCM's evidential mechanism involves the computation of summed similarities. For example, if the available category responses are A and B, then GCM calculates the sum of the similarities of the presented stimulus to each of the stored exemplars belonging to category A (S_A). The same calculation is performed for category B (S_B). Simplifying slightly (see below), GCM's decision mechanism is that the probability of a category A response is

$$\frac{S_A^\gamma}{S_A^\gamma + S_B^\gamma} \qquad \text{Equation 1}$$

, where γ was a subsequent addition to the model (Ashby & Maddox, 1993; Nosofsky & Zaki, 2002) in order to allow it to account for the degree of response determinism seen in participants. As γ becomes large, the probability of selecting the category with the larger summed similarity approaches 1 and the probability of selecting the category with the smaller summed similarity approaches 0. Equation 1 is typically expressed in a more general form that permits more than two response options (Nosofsky, 1984). The calculation of summed similarities also takes into account memory strength (represented as a multiplicative factor of the stored item's similarity score, Nosofsky, 1988) and the decision mechanism can include category response bias (represented as a multiplicative factor of the summed similarity, Nosofsky, 1984).

GCM has been the basis of a number of other models. The most influential of these is ALCOVE (Kruschke, 1992). ALCOVE is an instantiation of GCM within a localist

connectionist framework, and it provides a formalization of the process by which the selective attention and memory-strength parameters of GCM change over time. In ALCOVE, both sets of parameters are determined by gradient descent – an idea closely similar to the reduction-of-prediction-error accounts of learning and learned attention provided by animal learning models (Mackintosh, 1975; Rescorla & Wagner, 1972). Other extensions of GCM include the Extended Generalized Context Model (EGCM, Lamberts, 1995), which formalizes the assumption that stimulus representations are not perceived instantaneously, and the Exemplar-Based Random Walk model (Nosofsky & Palmeri, 1997a), which formalizes the assumption that categorical decision processes are not instantaneous.

In addition to the introduction of a highly influential formal model of categorization, work on GCM also demonstrated the potential of formal categorization models to provide very precise quantitative fits to observed phenomena. Indeed, the degree of precision that can sometimes be achieved by GCM is impressive (e.g. Nosofsky, 1986). The quantitative examination of a formal model can take a number of different forms and one approach concerns an emphasis on the minimization of an error term. As we will set out in a later section, evaluation of formal models *solely* on the degree to which they can minimize an error term can be problematic in a number of respects (as researchers working on GCM accept, see e.g. Nosofsky & Stanton, 2005, p. 613).

The GCM is the most widely known and understood of the current formal models of categorization. It also one of the most clearly specified. For these reasons alone, many of the examples employed in the current article are formulated in terms of the GCM. We do not intend to imply that the GCM is deficient compared to other models. In fact, we believe that the GCM's clear specification is a great strength, and one that has facilitated the writing of this article.

Nosofsky-Smith-Minda (NSM) prototype model

Prototype models assume that each category has a single intermediate representation. That intermediate representation – the category prototype – is typically considered to be the average of the representations of the category examples, although some models assume that distributional information is also stored (Fried & Holyoak, 1984). Formal prototype models of categorization can be traced back at least as far as Reed (1972), but arguably the most influential version of recent times has been the Nosofsky-Smith-Minda (NSM) prototype model, originally developed by Nosofsky (1987), and extensively investigated in the work of Smith and Minda (e.g. Smith & Minda, 1998).

One of the interesting properties of the NSM prototype model is that, except in the critical aspect of stimulus representation, it is closely similar to the GCM. For example, it employs the same similarity-distance equations (Figure 3D), and the same decision functions (Equation 1), as the GCM. Such similarity between models on all but one, theoretically interesting, issue facilitates model comparison. Work on the NSM model has included principled attempts to assess the relative adequacy of two qualitatively different formal models, for example the comparison of NSM with GCM (e.g. Smith, 2002). This is an approach to model comparison that we advocate throughout this article, as long as the comparison satisfies certain requirements (we will argue later that certain kinds of comparisons lead to more compelling conclusions than others). Comparison of qualitatively different models seems more likely to lead to progress in the field than evaluations of the fit of a single model or of the relative fit of a number of variants of an *a priori* favored class of model. We return to this point in more detail in a later section.

SUSTAIN

SUSTAIN (Love, Medin & Gureckis, 2004) is a formal model of categorization designed to account for both categorization probabilities and feature-inference probabilities. It also provides a formal model of the relationship between supervised and unsupervised category learning (i.e. category learning in the presence and absence of category labels), and makes different representational assumptions to either the GCM or the NSM prototype model.

SUSTAIN is able to provide an account of both categorization and feature inference as a result of being an auto-encoder – in other words, a model that seeks to reproduce and complete its input at its output. In such models (see also McClelland & Rumelhart, 1985) categorization and feature inference are differentiated by the nature of the information missing at input – in categorization, the category label is missing; in category-to-feature inference, the category label is present at input, but one or more of the features are absent. In both cases, the model takes this incomplete input, and attempts to reproduce it – with the missing information "filled in" – at its output.

In common with GCM and NSM, SUSTAIN represents stimulus input within a psychological space, and allows attentional modulation along the dimensions of that space. However, the attentional modulation in SUSTAIN affects the narrowness of the receptive field of cluster representations (see below), rather than GCM's uniform compression/ expansion of an entire dimension. SUSTAIN also incorporates a bias to focus on a subset of stimulus dimensions.

In terms of intermediate representations, SUSTAIN is neither an exemplar model, nor a prototype model. Instead, its representations are *clusters.* Exemplars and prototypes are special cases of cluster-based representation and, as is the case with exemplars or prototypes, clusters are represented as points in psychological space. In exemplar-based representation

there is exactly one cluster for each experimenter-defined stimulus; in prototype-based representation there is exactly one cluster for each experimenter-defined category.

SUSTAIN forms and develops clusters in a trial-by-trial manner. The first stimulus presented is assigned its own cluster, centered on that stimulus. In supervised categorization, subsequent stimuli are assigned to their own cluster if the existing clusters make an incorrect prediction about the category membership of the presented item. In unsupervised categorization, a stimulus is assigned a new cluster if it is sufficiently different to the existing clusters – how different it has to be in order to produce a new cluster is a free parameter (i.e. a parameter whose value is assumed to be whatever makes the model most accurate in predicting performance).

In addition to the recruitment of new clusters, SUSTAIN engages in a number of other types of adaptation. First, clusters compete to represent the input, and the "winning" cluster (the one most similar to the presented item) adapts by moving in psychological space towards the location of the presented item. Second, the winning cluster modulates dimensional attention in the direction that increases its activity. Third, where feedback is available, connections from clusters to output units change in accordance with a delta rule (Widrow & Hoff, 1960). As in ALCOVE, the basic intuition underlying this adaptation is that the model learns in order to reduce prediction errors. SUSTAIN does not formalize how connections from clusters to output units change in the absence of feedback (Love et al., 2004, p. 316).

COVIS

The COVIS model (Ashby et al., 1998) is unique in terms of the models considered here in that, from its inception, it has had both a computational and a neurological specification. The neurological specification of COVIS has motivated and guided some of the work on the neuroscience of categorization (for an overview see Ashby & Maddox, 2005).

The ultimate objective of the kind of approach exemplified by COVIS is that the computational and neuroscience components of a model should provide mutual constraints for each other. For example, the specification of intermediate representations should be constrained by the known neurophysiology of the systems that are hypothesized to support these representations. Equally, parameters in the computational part of the model can be related to neurological parameters - for example, COVIS links certain parameters in its learning equations to dopamine levels (Ashby, Paul, & Maddox, 2011).

COVIS has has three main components – an explicit system, a procedural-learning system, and a system that determines whether the explicit system or procedural-learning system controls responding. The intermediate representations of the explicit system in COVIS are unlike the other models so far discussed, in that the explicit system is seen as testing and selecting explicit rules about category membership. The set of rules considered by COVIS (the *candidate* rules) are one-dimensional (e.g. if length > X, then category A), and also sometimes includes rules constructed from one-dimensional rules in a Boolean manner (e.g. if length > X, **and** brightness > Y, then category A). For any one decision, only one rule controls the output of the explicit system – the *active* rule. If the decision is correct (as determined by feedback) then the active rule is unchanged. If the decision is incorrect, then a rule is selected from the set of candidate rules with a probability that reflects the rule's current *weight*. Rule *weight* is derived from rule *salience.* For active rules, salience increases with correct responses and decreases with incorrect responses (both changes are subject to some noise, however). The salience of inactive rules remains unchanged. Rule weight for the active rule is defined as its salience plus a constant representing the individual's tendency to perseverate. Rule weight for inactive rules equals their rule salience, with the exception of one randomly selected inactive rule, whose weight is increased by a mean of λ. The parameter λ represents the individual's tendency to switch rules. The outputs of the explicit

system are (a) a category decision (e.g. “category A”), and (b) a confidence score for that decision. The explicit system is considered to be supported by the prefrontal cortex, the anterior cingulate, and the head of the caudate nucleus.

The procedural-learning system[2] operates in a different way to the explicit system. As in the GCM and SUSTAIN, the input representation of the procedural-learning system is conceptualized as a psychological space. The intermediate representations in the procedural-learning system are different to GCM, NSM, and SUSTAIN. Rather than exemplars, prototypes, or adaptive clusters, the procedural-learning system assumes that the psychological space is covered by a large number of pre-existing, fixed, radial basis units. A radial basis unit is one whose output is maximal when the presented stimulus coincides with it in psychological space, but whose output drops rapidly as distance between the stimulus and the center of the radial-basis unit increases. In the procedural-learning system of COVIS, the output of radial-basis units drops off as a Gaussian function. One way of viewing this form of intermediate representation is as an exemplar model where a very large number of evenly distributed exemplar representations are assumed to exist, even when no exemplars have been seen.

Of course, under such circumstances, these representations contain no information about category membership. The procedural-learning system resembles ALCOVE in that it assumes information about category membership is contained in connections from the radial-basis units to response representations (an evidential mechanism). As in previous models we have discussed, these connections change in strength on the basis of feedback, with the principle of minimization of prediction error determining how these connections will change. The procedural learning system differs from these other models in that it assumes minimization of individual prediction errors (i.e. between a single radial-basis unit and a response unit; e.g. Mackintosh, 1975) rather than minimization of summed prediction error

(Rescorla & Wagner, 1972). The outputs of the procedural system are (a) a category decision, and (b) a confidence score for that decision. Note that, unlike the other models described in this article, the procedural-learning system of COVIS has no attentional mechanism. Effects attributed to selective attention in other models are the product of the low-dimensional rules typically employed by the COVIS explicit system. The neurological structures associated with the procedural system are the inferotemporal cortex and the tail of the caudate nucleus.

The outputs of the explicit system and the procedural-learning system both feed into a competition resolution system. This resolution system decides whether it is the explicit system or the procedural-learning system that controls responding on a given occasion. In deciding the winner of this competition, the resolution system takes into account two factors – the *trust* the resolution system has in each component system, and *confidence* each of the component systems have in their output. The system for which the product of confidence and trust is higher wins the competition. In COVIS, *trust* is a global value – the current trust value for the explicit system is θ_E, which ranges between 0 and 1, and the current trust value for the procedural-learning system is θ_I, which is constrained to be 1 - θ_E. Trust in the explicit system increases if its response is correct, and decreases if its response is incorrect. In typical applications, trust in the explicit system starts very high (e.g. 0.99).

In considering categorization responses to be the product of a competition between a rule-like and an exemplar-like process, COVIS formalizes a particular dual-system approach to categorization that can be traced back at least as far as Brooks (1978). Another, non-identical, formalization of a rules-and-exemplars theory is ATRIUM (Erickson & Kruschke, 1998). However, unlike ATRIUM, and unlike the other models discussed in the current article, the COVIS formulation is expressed in terms of the assumed underlying neuroanatomy and neurochemistry.

KRES

All the preceding models focus on situations where the participants' pre-experimental knowledge of the trained category structure is negligible. Whilst it is certainly easier to study categorization processes in the absence of any relevant pre-experimental knowledge, categorization outside the lab seldom operates in a knowledge vacuum. Indeed, the empirical study of prior knowledge effects on categorization has revealed a number of reliable phenomena (see Murphy, 2002, pp. 141-198 for a review). The Knowledge RESonance model (KRES, Rehder & Murphy, 2003) provides a formal account of some of these phenomena.

Input representations in KRES are different to those in any of the models so far considered. Stimulus dimensions are represented by a set of mutually exclusive and mutually inhibitory features. KRES also assumes that output representations inhibit each other - this use of mutually inhibitory output representations is analogous to the "pick the best" category decision rule of COVIS and is approximated by the choice rule used by GCM and NSM (although the approximation becomes poor with more than two categories; Wills, Reimers, Stewart, Suret and McLaren, 2000).

As in ALCOVE, and in the procedural-learning system of COVIS, category knowledge in KRES is represented by the formation of connections whose strength changes in accordance with the principle of reduction of prediction error. However, unlike ALCOVE or COVIS, KRES also permits connections between input units (see also McClelland & Rumelhart, 1985). In KRES, prior knowledge is represented in two ways – (a) pre-existing feature-feature connections, (b) pre-existing feature-category connections.

Another aspect that distinguishes KRES from models such as ALCOVE or COVIS is that KRES is a *recurrent* network. In all other models considered here, activation proceeds from input representations to output representations. In a recurrent network, activation also

proceeds from output representations to input representations, between different input representations and between different output representations. It is this resonance of information around the network that leads to some of the predictions of the KRES model concerning the effects of prior knowledge on categorization.

Simplicity model

The Simplicity model (Pothos & Chater, 2002) is a model of unsupervised categorization. It is the first model specifically developed to explain category intuitiveness, that is, to explain why certain classifications for a set of concurrently presented stimuli appear more natural to naïve observers than others. It assumes that preferred classifications will involve groupings that maximize within-category similarity and minimize between-category similarity, across all exemplars (see also Rosch & Mervis, 1975). Thus, like SUSTAIN, the simplicity model instantiates a preference for similarity-based groupings in unsupervised categorization. However SUSTAIN, unlike the simplicity model, has a bias towards groupings using a subset of the stimulus dimensions.

The simplicity model aims to predict the optimal number of categories in an unsupervised classification. It achieves this through a scheme for computing the codelengths for the similarity information between the items, with and without categories (the particular framework employed is Minimum Description Length, Rissanen, 1978). The codelength for similarity information with categories can be lower than the codelength without categories, if the categories can provide an efficient way of coding for this similarity information. Whether this is possible or not clearly depends on how categories can code for similarity information and the particular assumption in the simplicity model is that a category is a set of objects for which all within-category similarities are greater than any between-category similarity (following Rosch & Mervis, 1975). If the similarity structure of a set of objects is consistent with this definition (for a set of categories), then categories can simplify the description of the

corresponding similarity information. Note that the assumption of how categories code for similarity information is analogous to the specification of prior distributions in Bayesian approaches (cf. Chater, 1996).

The simplicity model assumes that the optimal number of categories appropriate for a set of objects is the number that reduces the codelength for describing similarity information for that set of objects the most. Also, the difference between the codelength with categories and the codelength without categories is a measure of the intuitiveness of the category structure. The latter is the unique contribution of the simplicity model, as no other model can immediately produce a value that can be interpreted as psychological intuitiveness (and indeed this has been a dependent variable neglected in categorization research). Having a quantitative measure of category intuitiveness can be very useful. For example, it allows the model to make parameter-free predictions about dimensional attention (Pothos & Close, 2008; cf. Colreavy and Lewandowsky, 2008).

The simplicity model's use of information theory comes at a price: the model has to assume a non-metric space, so that similarity information is represented in terms of relative magnitudes of similarities. This implies that, as long as categories are well separated, the degree of separation does not matter and also the spread of categories does not matter. These are important assumptions regarding the implementation of the simplicity model which have yet to be confirmed.

Rational model

The Rational Model of Categorization (RMC; Anderson, 1991) is a trial-by-trial model of categorization, based on Bayesian updating of probabilities. Specifically, it determines the classification of a novel instance in terms of how likely the instance's features are, given the observed features of the members of different categories. As a result, the RMC effectively favors clusters of items that are similar to each other. Like SUSTAIN, but unlike

the simplicity model, RMC has a free parameter that determines how dissimilar a new stimulus has to be to in order for it to form a new cluster. However, in SUSTAIN, this free parameter only applies to unsupervised categorization, whilst in the RMC, it applies to both supervised and unsupervised categorization. Also, like SUSTAIN, the RMC is able to provide an account of both categorization and feature inference. One way in which attentional selection can be implemented in the RMC is in terms of prior biases for particular dimensions (Anderson, 1991; for an alternative approach see Pothos & Bailey, 2009).

One aspect that the RMC shares with the GCM is that the RMC has formed the basis of a number of developments and related models. For example, Anderson and Matessa (1992) proposed a modification to account for people's sensitivity to feature correlations, and Sanborn, Griffiths and Navarro (2006) have proposed a variant that allows order-independent classification predictions.

Summary

The formal modeling of categorization is currently characterized by considerable diversity - these models differ on most aspects it would be possible for categorization models to differ. For example, the nature of intermediate representations (prototypes, exemplars, adaptive clusters, fixed radial basis units), the nature of selective attention, single vs. multiple systems approaches, feed forward vs. recurrent information flow, pick-the-best versus ratio rule (Equation 1) decisions, similarity-based versus Bayesian classification. Those aspects of the models for which there is consensus, or at least some convergence, tend to be constructs from outside categorization research, and about which the formal modeling of cognition as a whole has largely converged (e.g. adaptation as being driven by the minimization of prediction error – see Friston, 2010, for the wide applicability of this concept).

This high degree of divergence amongst formal models of categorization obviously poses a profound challenge to any experimental psychologist attempting to provide principled

empirical tests between different models. Moreover, it is hard to see these multiple models as complementary. In order for them to be complementary, there would have to be consensus on the situations in which each is best applied. This does not exist.

Reflecting on the arguments we made in favor of formal modeling, one might reasonably argue that formal modeling of categorization has led to an increased appreciation of the complexity of the problem, and also some deeper insight into empirical phenomena. However, the presence of multiple, domain-general, models subverts many of the other advantages of formal modeling – having multiple domain-general models does not serve the goals of ambiguity reduction or behavior prediction (except, of course, in the special case where all models behave in the same way). The way to rectify this problem is to make use of the other main advantage of formal models – their ability to facilitate theory comparison against empirical data. In the next section, we evaluate current practice in model comparison within the field of categorization research, and make a series of best-practice recommendations designed to maximize the chances for further progress. The issue of model comparison is clearly pertinent for many areas of psychology, including areas with a close relation to categorization such as recognition memory (Nosofsky & Stanton, 2006) and magnitude estimation (Bergert & Nosofsky, 2007); however, the specific examples upon which we draw in this article are from studies of categorization.

Model comparison

Model comparison, as defined here, is the comparison of at least two different classes of model that have some currency in the literature, where the comparison concerns the relative adequacy of those models to account for certain empirical phenomena. One example of this kind of model comparison is the work by Nosofsky and Stanton (2005). In that paper, the authors compared exemplar, prototype and decision-bound models on their ability to

account for the effects of probabilistic versus deterministic feedback on the accuracy and speed of categorization. An exemplar model provided the best account of these data.

In contrast, research that evaluates the ability of a single model to encompass certain phenomena does not constitute model comparison as defined here. For example, Nosofsky and Palmeri (1996) present a demonstration that the ALCOVE model can accommodate the results of a variant of the Shepard et al. (1961) experiment (see Figure 1) in which the stimulus dimensions are integral (rather than separable, as in the original demonstration). Such modeling work has considerable merit – it shows, for example, that there is at least one extant model that can account for what has been found. Nevertheless, work of this type seems unlikely to resolve the problem of multiple, divergent, non-complementary formal models of categorization, which is the focus of the current article.

Similarly, comparing variants of the same class of model is undoubtedly important in the development and refinement of a particular theoretical approach, but does little to solve the central problem of multiple, divergent, non-complementary models. For example, Nosofsky and Kruschke (1992) report (amongst other things) a comparison of the GCM model with a subsequent development of the GCM. Work of this type is useful in the sense that it helps motivate the development of models within a particular class, but does not directly address the problem of resolving relations between divergent, non-complementary, classes of model. A similar point pertains to comparisons where one model is well-established, but the comparison model has no currency in the literature, and the less-well established model is found to be inferior. Such comparisons have their uses, but they seem unlikely to resolve the problem we consider here.

There are numerous positive examples of model comparison in the categorization literature. For example, exemplar models have been compared against configural-cue models (Nosofsky, Kruschke & McKinley, 1992; Nosofsky et al., 1994), prototype models (Smith &

Minda, 2000), the Rational model (Nosofsky et al., 1994; Pothos & Bailey, 2009), and decision-bound models (McKinley & Nosofsky, 1996; Nosofsky & Palmeri, 1997b; Little, Nosofsky & Denton, 2011). And yet, limited progress appears to have been made in reducing the number of divergent, non-complementary models of categorization. Decision-bound models have been around in something approaching their current form for more than 20 years (Ashby & Gott, 1988), yet are still the subject of evaluation in current research (e.g. Little et al., 2011). Configural-cue models have also been a feature of categorization research for more than 20 years (e.g. Gluck & Bower, 1988) yet some of their key processing and representational assumptions live on in models such as KRES. Prototype models of categorization have been with us for at least 40 years (e.g. Reed, 1972), but still motivate current research (e.g. Homa, Hout, Milliken & Milliken, 2011). Why the apparent lack of confident progress towards reducing the number of divergent, non-complementary models of categorization?

One possibility is that, as these are all very complex models and as principled comparisons pose profound empirical, computational, and theoretical challenges, overall progress is inevitably slow. No doubt, this is part of the answer. Another possibility, and the one we explore in this article, is that progress is slower than it needs to because formal model comparisons in categorization have generally been rather narrow. For example, Smith and Minda (2000) presented an analysis comparing the GCM, the NSM prototype model, and variants thereof, against many replications of a study that examined response probabilities for a set of test items subsequent to training on one particular category structure (the "5-4" structure, introduced by Medin & Schaffer, 1978)[3]. Hence, the comparison was restricted not just to the same kind of evidence (classification probabilities) but effectively to variants of the same data set. Pothos and Bailey (2009) explored the ability of three different models (an unsupervised version of the GCM, the simplicity model, and the RMC) to account for five

different data sets. While initially promising, as it turned out, none of the models were clearly superior across all five data sets, showing that a low ratio of data sets to models (5:3) was not adequate to discriminate between these models (equally, that the particular data sets were non-diagnostic in this comparison).

There are numerous other examples where model comparison has been restricted to one or two experiments (e.g. Little et al., 2011; McKinley & Nosofsky, 1996; Nosofsky, Kruschke & McKinley, 1992; Nosofsky et al., 1994; Nosofsky & Palmeri, 1997b; Nosofsky & Stanton, 2005; Stanton, Nosofsky & Zaki, 2002). One might argue that narrow comparisons are the result of what can reasonably be achieved in a single research article. No doubt there is some truth in this argument, and researchers in the categorization field do appreciate the necessity for broader comparisons. However, narrow comparisons are not unavoidable in a general sense. For example, in the modeling of reading aloud, Perry, Ziegler and Zorzi (2007) compared three models against thirteen benchmark phenomena. In the final section of the current paper, we return to the issue of the extent to which broad comparisons are feasible.

We start from the, in principle, non-controversial point that a key goal for formal modeling must be to assess the relative adequacy of the numerous pre-existing models against a broader range of the known empirical phenomena - to not do so is to essentially negate most of the reasons for favoring formal models in the first place. In the current article, we consider the ways in which a formal model can be assessed against empirical phenomena and consider some of the reasons that have led to narrow model comparisons. Then, we identify the approaches in the literature that we consider to be the gold standard for model evaluation and development. We also list the range of DVs and IVs against which formal models of categorization could reasonably be expected to be assessed. Even though all these variables have been considered in the categorization literature (some more than others), we

think it is important to summarize them here, as in practice model comparison has been restricted to a handful of variables.

Assessing relative adequacy

Returning to our earlier definition, a formal model transforms changes in one or more IVs into changes in one or more DVs. If model X does this better than model Y, model X should be preferred over model Y – but how should relative adequacy be operationalized? Below, we make the case that relative adequacy should be assessed by comparing well-defined models on the basis of the number and proportion of irreversible, ordinal, penetrable successes in accounting for empirical phenomena. Each of the components of this operationalization of relative adequacy is discussed in the sections that follow.

Ordinal adequacy

One way to assess the empirical adequacy of formal categorization models is to evaluate their ability to minimize the quantitative difference between their outputs and some empirical observations. We describe this as *SSE adequacy* (SSE is an acronym for sum of squared errors, a common measure of quantitative difference). Assessing formal models solely on the basis of SSE adequacy has two serious problems:

(1) SSE does not distinguish between quantitative and qualitative adequacy, as illustrated in Figure 5. On an SSE measure, the two models in Figure 5 are indistinguishable – they have the same SSE. Yet, most theorists would agree that the model in Figure 5B provides a better account of the empirical results than the model in Figure 5A. This is because the model in Figure 5B correctly predicts that increases in the IV lead to increases in the DV, whilst Figure 5A makes the opposite prediction.

-- Figure 5 about here --

(2) A reliably lower SSE is not necessarily indicative of a more adequate model. Indeed, except in the purely theoretical case where measurement error is zero, the model with the lower SSE can sometimes be the *less* adequate model, if its lower SSE comes from its greater ability to fit noise. This phenomenon is described as *overfitting*.

Overfitting can be revealed by techniques such as cross-validation – one splits the data into a calibration and a validation sample (typically, the calibration and validation samples are two random subsets of the responses made by a participant). The model parameters are estimated via minimization of SSE on the calibration sample, and then the same parameters are applied to the validation sample. The greater the increase in SSE from the calibration sample to the validation sample, the more likely it is that the model overfitted the calibration sample.

Overfitting is a real possibility in the formal modeling of categorization. For example, Minda and Smith (2001) argued that a prototype model provided a better account of a particular set of data than an exemplar model on the basis of a small difference in quantitative fit (the prototype model was closer to the data by about three percentage points on average). In a replication that included cross-validation analysis, Olsson Wennerholm and Lyxzen (2004) demonstrated that the prototype model showed a greater increase in SSE from calibration to validation sample than did the exemplar model, with both models showing the same level of fit in the validation sample. This raises the possibility that the superior quantitative fit of the prototype model in the calibration sample (and, by extension, in Minda and Smith, 2001) was due to overfitting. Nosofsky and Zaki (2002) also queried the Minda and Smith's (2001) results, noting that where the GCM exemplar model included a response-scaling parameter (γ in Equation 1), it could accommodate Minda and Smith's results better than a prototype model. However, Olsson et al. (2004) demonstrated that the version of GCM including the response-scaling parameter also showed a greater increase in SSE from

calibration to validation sample, than did a version of GCM not including a response-scaling parameter (with both versions of GCM producing equivalent levels of fit in the validation sample). This again illustrates the potential for overfitting in the comparison of formal models of categorization. Of course, the issue of whether the inclusion of a gamma parameter leads to overfitting in the narrow comparison of three models to one experiment is different to the issue of whether the GCM (or any other model) requires a gamma parameter in order to be able to accommodate a broader range of results.

In summary, SSE is dissociated from important aspects of relative adequacy – two models can have the same SSE in cases where most theorists would agree one is superior, and better SSE can sometimes indicate a less adequate model. For both these reasons, we argue that the primary evaluation of formal models of categorization should be against a criterion of *ordinal adequacy*. In other words, we are suggesting that models should primarily be assessed first as to whether they capture the ordinal properties of a data set. For example, in the Shepard et al. (1961) data set (see Figure 1), this might mean getting the six problem types in the correct order of difficulty. Assessing adequacy by the ability to reproduce the ordinal properties of a data set eliminates the problem described in Figure 5 – the model in Figure 5B is the more adequate account under a criterion of ordinal adequacy. Adopting ordinal adequacy as the primary measure of success also reduces (but does not necessarily eliminate) the risks of illusory model superiority due to overfitting.

An ordinal adequacy criterion does not limit models to simple findings – one could assess, for example, whether a model could reproduce the ordering of the curvatures of a category acquisition function, or the kurtosis of a set of RT distributions. And making ordinal adequacy primary does not render SSE redundant. Where, across a broad range of phenomena, models are indistinguishable on an ordinal basis, cross-validated differences in

SSE provide a useful secondary measure of model adequacy (as do Bayesian methods of model selection, e.g. Boucher & Dienes, 2003; Pitt, Kim & Myung, 2003).

There are cases in the literature that include assessments of ordinal adequacy, including many of those we referred to in the earlier discussion of our definition of model comparison. For example, in Nosofsky et al. (1994), GCM is shown to make an ordinally different prediction to certain configural-cue models (Gluck & Bower, 1988), with the data being consistent with the GCM's predictions. Similarly, in Nosofsky and Palmeri (1997b), the EBRW model is shown to make an ordinally different prediction to a decision-bound model, with the data being consistent with the EBRW model.

There are also cases where model comparison has proceeded solely on the closeness of quantitative fit, with both models being able to accommodate the ordinal pattern observed. For example, McKinley and Nosofsky (1995) concluded in favor of a variant of the GCM model (over a decision-bound model) solely on the basis of degree of quantitative fit. Similarly, Shin and Nosofsky (1992) report a comparison in which GCM accounted for 98% of the variance whilst the prototype model accounted for 94% of the variance.

In summary, some model comparisons in categorization research have included ordinal success as part of their evaluation, whilst others have relied solely on the quantitative closeness of fit. Our argument is that comparisons that include a consideration of ordinal success represent best practice, for the reasons outlined above. Of course, regardless of whether evaluations are based solely on SSE, or whether they additionally include consideration of ordinal success, it remains important that models provide an account of as much as the collected data as possible, rather than focusing on one or two collected data points at the expense of ignoring the rest of the data.

Functions of quantitative adequacy

To clarify, we are *not* advocating a complete avoidance of a quantitative approach to model evaluation. We believe quantitative adequacy, when considered in combination with ordinal success, can serve important functions. For example, whilst the focus of the current article is on formal model comparison, this is not the only way in which formal models can be employed. Another use of formal models is as an existence proof that a particular model has the potential to encompass a particular result. One example of this approach is Nosofsky and Zaki's (1998) demonstration that a version of GCM can account for the fact that amnesics are sometimes more impaired on old-new recognition than they are on categorization (Knowlton & Squire, 1993), a result previously considered to be outside the scope of single-system theories. The impact of such existence proofs seems to be increased if the formal model captures not only the ordinal patterns of the experiment, but also provides a striking degree of quantitative closeness. The issue of what degree of closeness is required to be sufficiently impressive is, of course, rather vague in situations where only one formal model is considered. Nevertheless, it is beyond dispute that quantitatively close existence proofs can have a profound impact on the field (as measured by, for example, the number of citations they receive).

Another potential use for quantitative adequacy is in situations where all models under comparison capture the ordinal patterns in the data. Under such circumstances, one may wish to favor the model that produces the closest overall quantitative fit. In situations where one is confident that the difference in quantitative fit does not result from overfitting (see above), closeness of quantitative fit may provide some useful additional information, both in terms of relative model success and in terms of estimation of parameter values (as parameter values can provide information about how models account for an empirical finding). More generally, ordinal success is, by definition, a coarser measure than closeness

of quantitative fit – the latter takes into account the magnitude of effects whilst the former does not. It is conceivable that, as the comparison of models against a broad set of data proceeds, a trade-off will emerge where Model X accounts for more ordinal patterns than Model Y, but at the expense of having lower quantitative adequacy than Model Y. The issue of which model is the more adequate under these conditions would rightly be a topic for serious debate, and a measure of quantitative adequacy would clearly be necessary to inform that debate.

Irreversible success

We argued earlier that one of the main advantages of formal models over more informal forms of theorizing was the potential of formal models for ambiguity reduction. We also argued that one reason this potential had failed to be realized in the formal modeling of categorization was the presence of multiple domain-general models and no consensus on the relative adequacy of these models. Here, we emphasize that achieving progress towards consensus requires an avoidance of *arbitrarily variable parameters*, and an evaluation of the relative adequacy of models through an examination of the *irreversible successes* that can be attributed to them. Below, we provide a definition of the concept of arbitrarily variable parameters, illustrate why they are a problem, and propose the assessment of relative adequacy through irreversible modeling successes.

Arbitrarily variable parameters

A model parameter is some (usually numerical) information that is part of the model specification, rather than provided via the IV inputs. Most models have parameters, including some of the most successful and elegant formal models ever created (e.g. Newtonian gravity). Having parameters, even a large number of parameters, does not in itself cause any problems of ambiguity – it is what the modeler does with those parameters that is at issue.

In the formal modeling of cognition, the term *free parameter* is in common usage. We define a free parameter as any parameter whose value is determined as part of the process of determining model adequacy. Determining optimal values for free parameters can be seen as part of the process of model development, and the presence of free parameters has no necessary consequences for model ambiguity – as long as the values of those parameters are *universal*. Universal free parameters are those whose specification is general to the whole domain of phenomena that the model is intended to address. By contrast, an *arbitrarily variable parameter* is one that can take different values for different levels of an independent variable, and where each of those values is determined through a process of maximizing model adequacy (as opposed to, for example, being determined by independently measurable properties of the stimulus, environment or participant).

The problem of arbitrarily variable parameters

Allowing parameters to take freely determined values for different levels of an IV can cause severe ambiguity if changes in the value of that parameter are able to cause ordinal changes in the model's output. For example, Wills, Suret and McLaren (2004) examined whether pre-exposure to two different stimulus types facilitated or retarded subsequent categorization of those stimuli. Let's consider the IV here to be stimulus type (two levels – noise distorted vs. re-arrangement distorted) and the DV to be the direction of the exposure effect (two levels – retardation or facilitation). On that basis, four ordinally different things could have happened – of course, only one actually did (noise-distorted stimuli were facilitated; re-arrangement-distorted stimuli were retarded). One approach to modeling this experiment with the GCM would have been to allow *c* (the parameter controlling the overall expansion of psychological space, see Figure $C) to take four different values, one for pre-exposed noise-distorted stimuli, one for non-preexposed rearrangement-distorted stimuli, and so on. By selecting four appropriate values for *c*, the results of Wills et al. (2004) could be

accommodated by the GCM. But so could the three other possible results of this experiment that were *not* found. Indeed, the use of arbitrarily variable parameters in this case leads the GCM to become what is described as a *degenerate* model (Smith, Chapman & Redford, 2010)[4].

An alternative approach to modeling the results of Wills et al. (2004) with the GCM would be to use as input to the GCM psychological spaces derived from similarity ratings taken both before and after pre-exposure. This might capture the representational changes that result from exposure, and might have allowed the GCM to fit the data with a single set of parameters for all four conditions. Such an approach does not provide an account of representational change, but it (a) provides a clear statement that the form of representational change observed in Wills et al. (2004) is outside the explanatory scope of the GCM, and (b) removes arbitrarily variable parameters from the model specification in this context. For both these reasons, this second application of GCM is more useful in assessing model adequacy than the first application[5].

Defining irreversible success

The second application of GCM, if it worked, would also be an example of a model without arbitrarily variable parameters – but only in the microworld of the experiment discussed. Absence of arbitrarily variable parameters must properly be defined across the entirety of the data sets to which a model is applied – not just the context of a single study.

An ordinal success in reproducing the effects of IVs on DVs, in the absence of arbitrarily variable parameters, is what we describe as an *irreversible* success. The success is irreversible in the sense that turning one particular success into a failure (or, perhaps more appositely, a failure into a success) cannot be done without re-evaluating the model's ability to fit the entire data set that defines the model's domain. Derivation of a model's parameters with respect to the full canon of data that represents its domain, rather than independent

derivation of these parameters for each experiment (or even each condition of each experiment), ensures the model's successes are irreversible in the sense we have defined it here.

Number of empirical successes

A model that accommodates more of what we know empirically is, other things being equal (see later sections), a better model. Hence, our proposal is that relative adequacy of formal models can be assessed on the basis of the number of irreversible ordinal successes that can be attributed to them. This proposal contrasts sharply with current practice in categorization research, which is to examine in depth the results of a single or a handful of experiments, rather than seek breadth. For example, the original publication of GCM (Nosofsky, 1984) assessed the model against the result of just one, at that point unreplicated, study with six participants (Shepard et al., 1961). Twenty years later, the original publication of SUSTAIN (Love et al., 2004) assessed the model against seven experiments. One commendable aspect of the original assessment of SUSTAIN was that it employed universal free parameters – in other words, parameters that had a common value across all seven studies. SUSTAIN therefore achieved 7 irreversible ordinal successes in its original publication.

Both GCM and SUSTAIN have subsequently been assessed against other data. However, in both cases, these assessments have largely been performed independently of the original assessments. In other words, subsequent publications have determined the value of the model's parameters on the basis of maximizing the model's ability to reproduce the results of the particular studies considered in that paper. Against the criteria we are proposing, these additional publications do not necessarily demonstrate an increase in the number of irreversible ordinal successes of the model, and therefore do not necessarily reflect incremental progress on this criterion. Note that, in employing GCM and SUSTAIN as

examples, we do not intend to imply that these problems are specific to those models, or even that they are specific to models of categorization (for a related argument in the formal modeling of perception, see Pitt et al., 2003, p. 30).

One way to have met the proposed criteria would have been to determine the values of free parameters by their ability to maximize the number of ordinal successes across the combined data set – in other words, all studies against which the model had previously been compared, plus the additional data being considered in the new paper. Of course, it may be the case that different data sets will require different values for the model's parameters. As previously stated, a model with a large number of parameters is not necessarily ambiguous – what matters is whether those parameters are arbitrarily variable. For example, the attentional parameters in Nosofsky's (1984) fit of the GCM to the data of Shepard et al. (1961) are not arbitrarily variable because they are constrained by the hypothesis that dimensional attention is allocated to maximize categorization accuracy (a hypothesis subsequently given a formal mechanism in the ALCOVE model). This hypothesis results in the attentional parameters of GCM taking different values for the different conditions of Shepard et al. (1961). However, this variation is not arbitrary – in fact, it means that there are essentially zero free parameters for attention in that application of GCM.

As an illustration of the shortcomings of evaluating results in isolation, consider the work of Medin and Schaffer (1978). In one of the most influential results in the early development of exemplar theories, Medin and Schaffer demonstrated that, within the category structure shown in Figure 6, participants learned to respond correctly to stimulus A2 more quickly than they learned to respond correctly to stimulus A1. This occurred despite the fact that the features of A1 are in some sense more typical of Category A members than are the features of A2. Note that properties denoted "1" in Figure 6 are characteristic of Category A (occur more often in Category A than Category B). Hence A1 contains 3 properties

characteristic of A, whilst A2 contains 2 properties characteristic of A – A1 is more typical, but A2 is learned more quickly. Prototype models cannot accommodate this result, but exemplar theories can.

What is perhaps not always appreciated about the success of exemplar models in this context is that it is parameter dependent. For example, as Smith and Minda (2000, Figure 8) illustrate, both the observed ordinal pattern, and its opposite, can be accommodated by the GCM. One reaction to this state of affairs is to point out that GCM's prediction of an A2 advantage is robust across a broad range of parameters, and that the few parameter values that lead to the opposite prediction also lead to other, falsifiable, predictions (Nosofsky, personal communication). As a matter of logic, therefore, the question of what the GCM predicts about this seminal result cannot be resolved by considering this one result in isolation. It can only be resolved through the evaluation of GCM against a broad range of phenomena in a manner that avoids arbitrarily variable parameters.

One practical problem of our proposed "more is better" approach is the sheer size of the known phenomena. To give an idea of the scale of the problem, the number of experiments to which SUSTAIN was originally fit is lower than the number of publications on the subject of categorization in a single year of a major journal, and the number of publications since 1961 on the topic of categorization exceeds 2,000[6]. Of course, it's good practice to only model phenomena that have been independently replicated. Also, many studies are superseded by later work, and some provide insufficient information to allow any of the candidate models to be assessed against them. Despite all this, the number of reliable phenomena is undoubtedly large compared to the number of phenomena against which categorization models have currently been compared. And yet, it remains our contention that formal models of categorization should address them all. One key to making this a practical proposition, we believe, is to compare the relative adequacy of models that are *well-defined*

in terms of their explanatory scope. The importance of ensuring models are well-defined is discussed in the next section; below we consider an alternative to counting successes.

Elimination by failure

One possible alternative to the approach of counting successes is the approach of the elimination of formal models by the detection of their parameter-independent failures. For example, Nosofsky and Stanton (2005) argued that a particular decision-bound model, irrespective of its parameter settings, is unable to predict any error or reaction time difference between the deterministic and probabilistic feedback conditions of their experiments. As a difference is found in those experiments, the decision-bound model clearly cannot accommodate the results found, and hence a parameter-independent failure of that model is demonstrated. Similarly, Nosofsky and Zaki (2002) demonstrated that the best fit of a prototype model to a particular data set produced ordinally incorrect predictions about the observed classification of exception items. The fact that a model's best fit produces an ordinal failure implies that the failure is parameter independent.

The question of whether progress is best served by consideration of the numerosity of successes, or by specific falsifications, is as much a philosophical as an empirical question. Nevertheless, we tend to be persuaded by Kuhn's (1962) historical analysis that specific falsifications do not, in practice, lead to the rejection of theories. We believe this is also the case in categorization research - one illustration is that, more than 30 years after the publication of Medin and Schaffer's (1978) falsification, prototype-based accounts of categorization still seem to be the subject of research (e.g. Homa et al., 2011). Specific falsifications tend to lead, not to the rejection of theories, but to either a re-definition of the scope of the model that excludes the problematic result, or a refinement of the theory that turns failure into success. For example, whilst Medin and Schaffer's (1978) result is robust within the design they employed (see e.g. Minda & Smith, 2002), prototype theorists have

accommodated this, and other apparent problems for prototype theory, in a variety of ways, including a proposal that there is a transition from prototype representations to exemplar representations during the course of training (Smith & Minda, 1998; but see also Johanssen & Palmeri, 2002).

In some cases, responding to falsifications by re-definitions and modifications may constitute incremental progress. In other cases, it may result in a set of mutually incompatible post-hoc changes. As with a success-based approach, the only way to evaluate the relative (in)adequacy of different models is to consider them against a broad range of data in such a way that one avoids the problem of arbitrary variation. Although the approach advocated in the current article focuses on successes rather than failures, many of the issues remain the same whichever of these two perspectives is adopted.

Well-defined formal models

A well-defined model of categorization is one that considers all IV-DV combinations appearing in peer-reviewed publications on the topic of categorization. Well-defined models must either provide an account for an IV-DV combination or declare that the DV, IV, or both are outside its scope. For example, model X might provide an account of the effects of the duration of stimulus presentation on categorization accuracy. Model Y, in contrast, might not seek to provide such an account. Our argument is that Model Y should still contain a clear statement that this effect is beyond its explanatory scope.

It should be a straightforward matter to propose a model that is well-defined in terms of DVs – the modeler simply has to identify the DVs that are <u>included</u> in the model's explanatory scope. Making a model well-specified in terms of IVs is harder, because our definition of "well-defined" requires the modeler to make a statement about IVs outside the model's explanatory scope (e.g. the model does not account for changes in stimulus presentation time, and is applicable to self-paced situations). The large number of IVs makes

it likely that, in practice, being well-defined with respect to IVs will be a matter of degree. In a later section, we suggest a list of several major categories of IV against which it would be straightforward to define a model's domain. This would represent a substantial improvement on current practice, where formal models of categorization tend to be ill-defined with respect to all IVs for which they do not explicitly define a transformation.

A second advantage of ensuring models are well-defined is that it opens up the possibility of a secondary measure of relative adequacy – proportion of successes. Where two well-defined models have a comparable number of successes overall, we argue one should favor the model with the *smaller* explanatory scope (and hence the higher proportion of successes). For example, consider a model ("Model X") that accounts for all of the results in its defined domain ("Domain A"). Such a model is adequate and useful in the sense that, within Domain A, it provides a good model of behavior. By contrast, if Model Y accounts for only 10% of results in Domain A (and so neglects to provide an account for 90%), it provides a less adequate model of Domain A, regardless of the fact that Model Y might be providing partial coverage of several other domains (so that *overall* the number of Model Y successes are comparable to the number of Model X successes). Providing an adequate model of a well-defined but narrow domain seems to be of more use (in, for example, the sense of ambiguity reduction, behavior prediction and automated cognition) than providing a model that neglects to provide an account for the majority of the data across a broad domain.

Penetrable models

By the term *penetrable* we intend to denote two different but related things – the effort required to apply a formal model, and the extent to which the verbal description that must inevitably accompany a formal model allows some understanding of the model's processes in psychological terms. A more penetrable model is, other things being equal, a

better model. Below, we consider the *effort* and *psychological explanation* components of our definition of a penetrable model.

Effort

People often apply simple, incorrect models in preference to complex correct ones (e.g. Newtonian gravity instead of general relativity). Presumably, the reason they do this is that there is a real cost, in terms of time and effort, in understanding and applying the more complex model. Using the simpler, incorrect model is rational to the extent that the error introduced by the simpler model is minor compared to the additional effort required to apply the more complex, correct model. On this basis, we propose that where two models are approximately comparable in terms of the number and proportion of ordinal irreversible successes, the model that requires less effort to understand and apply is the relatively more adequate model.

Of course, the effort required to understand and apply a model is context dependent. For example, theorists with a background in connectionist modeling are likely to find a new connectionist model easier to understand and apply than a model whose basic formulation they are less familiar with (a model based on Bayesian rational analysis, perhaps). The comparison of models with different underlying bases might be encouraged by actions that reduce the effort required to understand and apply them. Two such actions are the provision of model summaries designed to be accessible to a more general audience (see, for example, Pothos & Wills, 2011), and the provision of publicly accessible computer-based implementations of models. Ideally, such implementations should be both transparent (for example, with relevant computer programs freely distributed as source code in a high-level computing language) and user-friendly (for example, by the provision of an associated user-friendly interface).

Psychological explanation

It is our contention that one property of a good formal model is that its formally-expressed components are accompanied by a natural language approximation. In some cases, this approximation may be a modification to the meaning of, or a disambiguation of, an everyday term, or a pre-existing verbal psychological construct. There are at least five reasons why good models should have a natural language approximation. First, a natural language approximation permits approximate understanding of the formal model with lower effort. For example, the GCM and the NSM prototype model make assumptions about category representation that have clear natural language approximations (exemplars, prototypes). Second, a natural language approximation will sometimes help clarify which aspects of a formal model are strong theoretical commitments, and which are not. For example, the natural language approximation for COVIS makes it clear the model has a strong theoretical commitment to the idea that categorization results from the competition between multiple processes. This, in turn, enables empirical tests which are better focused to the core assumptions in a model.

A third reason why good models should have a natural language approximation is that, by so doing, the model makes more connections to the bulk of psychological theorizing – which, of course, remains informal. For example, in GCM, *w* - the multiplicative factor by which distance on a psychological dimension is modulated - has the natural language approximation "attention weight". The presence of this natural language approximation allows one to make predictions about the relationship between *w* and certain DVs that would have been impossible otherwise - for example, relationships between *w* and eye gaze duration (Rehder & Hoffman, 2005).

A fourth way in which a natural language approximation helps to make connections with the bulk of psychological theorizing is by clarifying the central organizing principle of

the model. For example, the simplicity model is based on the central organizing principle that cognition can be understood as a process of re-encoding information in order to reduce redundancy.

A fifth situation in which natural language approximations are important is where formal models are used as a form of data-reduction technique. For example, Johansen and Palmeri (2002) employed the ALCOVE model, and a single-dimension variant thereof, as a data-reduction analysis in support of the claim that categorization strategies change from unidimensional to multidimensional with increasing experience of the stimuli. The ability to provide a meaningful natural language approximation of the differences between models affects the theoretical impact of such techniques.

A potential concern about natural language approximations is that, if applied in an overly-rigid manner, they might impede the development of truly novel explanations (as truly novel concepts, by definition, would not have a single pre-existing word that encapsulated them). One response to this concern is that natural language approximation should properly be an interactive process, with the link between a formal model and a natural language term leading, over time, to changes in the interpretation of that natural language term. For example, the etymology of the word *mass* can be traced back to the 11th century but the meaning of this word underwent substantial development as a result of Newton's work in the 17th century ("mass, n.2", 2011). In psychology, the terms "rule" and "similarity" currently have a number of different definitions (Pothos, 2005) - it may be that the contextualization of such terms via reference to specific formal models may lead to a disambiguation of the sense a user intends. In more extreme cases, one might have to invent new verbal labels for a truly novel insight (e.g. "quark", which was invented in the 1960s to represent developments in quantum mechanics; "quark, n.2", 2011).

Summary

In this section, we have argued that the relative adequacy of formal models should be assessed in terms of the number and proportion of irreversible ordinal successes produced by well-defined penetrable models. Although most formal models of categorization are relatively penetrable on our above definition, progress to date against other aspects of our definition of relative adequacy has been limited. Model comparisons typically involve assessments of a small subset of current models against a narrow data set. Even leaving aside the issue of *relative* adequacy, the assessment of the number of irreversible successes attributable to any particular model has been limited because models are typically fitted to one experiment (or even one condition of one experiment) at a time, with the values of the model parameters being determined independently in each case. A collection of independently derived model fits does not constitute incremental progress against a criterion of number of irreversible successes, as it leaves uncertain the extent to which the model's apparent success across the set is due to the presence of arbitrarily variable parameters. Number of irreversible ordinal successes is increased by demonstrating that the model can simultaneously produce those successes across the multiple data sets.

Another problem with the current state of formal modeling of categorization is that most models are largely ill-defined with respect to IVs. In the next section, we suggest several IVs against which all formal models of categorization could reasonably be expected to be well-defined.

Priorities and explanatory scope

As noted earlier, the number of empirical categorization phenomena, and hence the number of DVs and IVs for which a formal model might hope to account, is large. Two ways to make manageable progress in the evaluation of formal models of categorization against empirical phenomena would be to (a) define priorities with respect to which DVs are the

most important, and (b) agree a set of IVs against which a model's domain could be defined. The following two sections make specific proposals regarding these two issues.

Prioritizing DVs

The primary DVs for a formal model of categorization are those that directly index the classification decisions that are made. For example, accounting for patterns in the probability of a particular classification response must take priority over, for example, accounting for patterns in eye-gaze dwell time (e.g. Rehder & Hoffman, 2005). To illustrate, consider a situation where Model X accommodates the classification probabilities but not the eye-gaze measure, whilst Model Y accommodates the eye-gaze measure but makes ordinally incorrect predictions about the classification probabilities. We contend that, as a model of categorization, Model X is superior in this instance, as the primary purpose of a model of categorization behavior must be to account for classification decisions. It follows from this contention that, in a model of categorization, correctly accounting for classification probabilities has a higher priority than correctly accounting for eye-gaze durations.

We describe DVs that directly index the outcome of categorization decisions as primary DVs, and other measurements as secondary DVs. On the basis of that definition, the measurement of event-related potentials (e.g. Wills, Lavric, Croft & Hodgson, 2007) is another example of a secondary DV, whilst measures of classification consistency (e.g. Haslam et al., 2007) or, relatedly, intuitiveness (Pothos & Bailey, 2009) are other examples of primary DVs on this definition. The question of whether reaction time measures (e.g. Nosofsky & Palmeri, 1997a; Wills & McLaren, 1997) are primary or secondary is perhaps more controversial. Applying the criterion above, we argue that a model that can account for classification probabilities but not reaction times would, we believe, be generally considered to be more successful than a model that could account for reaction times but not classification

probabilities. On that basis, reaction time is a secondary DV, although we suspect that this assertion would not be universally accepted.

Secondary DVs nevertheless serve two important functions. First, they enable one to select between models with comparable numbers of successes on primary DVs. Second, they seem likely to serve a vital function in model development. Categorization models have internal states and it would be odd to ignore data that is potentially informative about such internal states. However, the relative adequacy of formal models of categorization should be assessed primarily against successes for primary DVs.

Defining domains through IVs

Below are a range of IVs (more accurately, a range of sets of IVs, each set corresponding to a different source of empirical evidence) against which any formal model of categorization might reasonably be expected to define its domain. The list is not exhaustive - other important IV sets, not covered here for brevity, include cultural differences (Medin & Atran, 2004), personality variables (Ward, 1983; Wills, Longmore & Milton, 2011), and other individual differences (DeCaro, Thomas & Beilock, 2008; Milton, Longmore & Wills, 2008; Tharp & Pickering, 2009).

Time pressure - There are a number of well-documented effects of time pressure on the acquisition of categories and on categorical decisions (e.g. Lamberts & Freeman, 1999; Milton et al., 2008; Smith & Kemler Nelson, 1984). The EGCM development of GCM provides an account of a subset of these phenomena; Ashby et al. (1998) make an informal argument that COVIS can account for a different subset. The remaining models neither provide an account nor define their domain with respect to this set of IVs.

Incidental categorization and concurrent load – There is a rich data set on situations where categorization occurs in the context of another task and how these situations can qualitatively affect categorizations decisions compared to situations of full attention (e.g.

Brooks, Squire-Graydon & Wood, 2007; Kemler Nelson, 1984; Milton et al., 2008; Smith & Kemler Nelson, 1984; Waldron & Ashby, 2001; Wills et al., 2011; Zeithamova & Maddox, 2006). ALCOVE has been applied to a limited subset of these data (Nosofsky & Kruschke, 2002), COVIS to a slightly larger subset (e.g. Ashby et al., 1998; Waldron & Ashby, 2001; Zeithamova & Maddox, 2006; but see also Newell, Dunn & Kalish, 2010).

Learning – DVs such as categorization accuracy and consistency demonstrably change with increasing experience of the stimulus set in both supervised and unsupervised categorization (e.g. Nosofsky, Gluck, Palmeri, McKinley & Gauthier, 1994; Wills & McLaren, 1998). Learning is occasionally accounted for by the GCM and NSM, but typically in an unconstrained way (e.g. Smith, Chapman & Roberts, 2010) or by assuming a linear relationship between frequency and memory strength (e.g. Nosofsky, 1988) – an essentially Hebbian assumption that empirical work indicates is incorrect (Griffiths & Mitchell, 2008). SUSTAIN and ALCOVE provide more constrained, prediction-error-based accounts of acquisition (e.g. Love et al., 2004; Nosofsky et al., 1994), and the Rational Model has also been fit to acquisition data (e.g., Nosofsky et al., 1994). The other models discussed in the current article are ill-defined with respect to this IV set.

Background knowledge – Background knowledge effects have been extensively examined in categorization (Murphy, 2002). KRES has been fitted to a subset of these data. It is unclear whether other formal models should be considered to include background knowledge effects as part of their domain.

Representational development – Many real-world stimuli are, compared to those typically used in categorization experiments, highly complex but also highly familiar. Perceptual difficulty, and level of familiarity with complex stimuli, can qualitatively affect the type of categories formed (Wills & McLaren, 1998) as well as classification accuracy (Attneave, 1957; McLaren, Leevers & Mackintosh, 1994; Wills et al., 2004). As discussed

previously, one approach to representational development in the GCM is to accommodate it (e.g. via post-categorization multidimensional scaling, Palmeri & Nosofsky, 2001), rather than model it as such. This is a fairly clear, albeit implicit, example of a model defining its domain as excluding a set of IVs. The domain of the other models discussed in the current article, with respect to IVs related to representational development, is less clear.

Number of categories – Most models of categorization are able to accommodate any number of category response options, and most implicitly provide an account of the effects of manipulating this IV as a by-product of adopting a variant of the Luce choice axiom (Equation 1). However, there is evidence that the Luce choice axiom does not correctly predict the relationship between 2-choice and 3-choice categorization probabilities (Wills et al., 2000). It is unclear whether models that employ the Luce choice axiom would therefore wish to define their domain as being restricted to situations where there are just two response options.

Unsupervised categorization – People are able to form category representations in the absence of feedback (Fried & Holyoak, 1984; Homa & Cultice, 1984). SUSTAIN, the Rational model, and the Simplicity model specifically allow for this possibility, and Pothos and Bailey (2009) have shown that the GCM can be modified to make predictions in this domain. Hence, many of the models discussed in the current article seem to include unsupervised categorization within their domain.

Abnormal brain function – In order to accommodate the effects of brain damage, formal models have to, at a minimum, clearly specify how that damage affects the parameters of the model. The GCM has been applied in this way to some relevant data on categorization and brain damage (Nosofsky & Zaki, 1998). The COVIS model makes specific predictions regarding how certain kinds of brain damage affect the categorization process (Ashby & Ell, 2001; Pothos & Wood, 2009) and, subsequent to its original publication, the SUSTAIN

model has been given a neuropsychological interpretation (Love & Gureckis, 2007). It is less clear for the other formal models discussed in this article whether they are intended to exclude abnormal brain function from their domain.

Developmental differences – A rich data set demonstrates that children sometimes produce qualitatively different classification decisions to adults (e.g. Kemler, 1983; Minda, Desroches & Church, 2008). COVIS (see Ashby et al., 1998) and a variant of GCM (Smith, 1989) has been applied to a subset of these data, SUSTAIN to a different subset (Gureckis & Love, 2004). For the other models reviewed in the current article, it remains unclear whether they should be considered to be specific to adult categorization behavior.

Species differences – Humans are not the only species able to categorize – impressive categorization abilities have also been reported in non-human primates (e.g. Delorme, Richard & Fabre-Thorpe, 2000), and in pigeons (e.g. Herrnstein & Loveland, 1964). From an evolutionary standpoint, it seems unlikely that the categorization processes of different species are entirely unrelated – and this is underlined both by reports of striking correspondences between human and non-human categorization performance in some cases (e.g. Smith, Chapman & Roberts, 2010; Smith, Beran, Crossley, Bloomer & Ashby, 2010; Wills et al., 2009) and by the extent to which formal models of human categorization have closely related antecedents amongst the formal models of animal learning. For example, Pearce's configural model (1987, 1994) is similar to ALCOVE in some respects; EXIT (Kruschke, 2001) can be considered as a generalized version of Mackintosh's (1975) theory of learned attention; SUSTAIN bears some resemblance to Saksida's model of perceptual learning (Saksida, 1999). GCM, COVIS, and NSM have all been applied to a subset of the data on categorization in non-humans (Smith, Chapman & Roberts, 2010; Smith et al., 2010). The remaining models are, presumably, intended to be human-specific.

Summary

The process of assessing formal models of categorization against the very large set of empirical categorization phenomena can be made more manageable by (a) prioritizing phenomena where the DVs are measures of decisions (as opposed to, for example, ERP components), and (b) ensuring models have well-defined domains with respect to IVs such as time pressure, full vs. divided attention, category acquisition, background knowledge, representational development, number of response options, supervised vs. unsupervised categorization, abnormal brain function, age, species, cultural background and personality variables. At present, formal models of categorization seldom define their domain for variables other than those for which they specifically seek to provide an account.

Model Comparison and Neuroscience

The models considered in this article differ dramatically in the extent to which neuroscience informed their formulation. At one end of the continuum, the Rational model and the Simplicity model formalize certain assumptions about the adaptive function of categorization, without reference to the neurological substrate that must necessarily instantiate these functions. Both the Rational Model and the Simplicity model, as currently implemented, involve computations which are intractable with moderately large numbers of items (but see Sanborn et al., 2010). For example, in the Rational Model, computation of classification probabilities involves a normalizing term across all possible classifications. Even for as few as 16 stimuli, this term would involve more than 10 billion computations.

GCM and the NSM prototype model make specific process-level assumptions about categorization, but these assumptions exist independently from any neurological mechanism that might instantiate them. ALCOVE, SUSTAIN and KRES are somewhat informed by neuroscience, in the sense that they are connectionist and hence their microstructure has been

partly informed by what is known at an individual neuron level. Indeed, neural plausibility at this level is one of the defining properties of connectionist modeling. However, whilst ALCOVE and KRES are somewhat neuron-like at their individual unit level, the architecture of ALCOVE and KRES seems to have been developed independently from information at the level of gross neuroanatomy. SUSTAIN is a slightly different case in the sense that although gross neuroanatomy was not discussed originally (Love et al., 2004), a consideration of the model in these terms did subsequently occur (Love and Gureckis, 2007)

As previously discussed, COVIS is unique amongst the models considered in that its conceptualization from the outset was strongly informed by hypotheses about how different brain regions might compete and co-operate in the classification of objects. Interestingly, COVIS departs from connectionist modeling, in that parts of the computation of the explicit system are symbolic and the manner in which this symbolic system is implemented at the level of single neural-like units is not as yet specified (this is not to say that such an implementation is impossible, simply that no specific implementation forms part of the COVIS model at the current time).

Throughout this article, we have emphasized the importance of comparing models over as broad a set of phenomena as possible. Specifying models in both computational and neuroscience terms will likely allow them to cover a broader range phenomena, in the simple sense that the neuroscience specification is an additional source of model constraints (and so of model predictions for behavior). Of course, it is not the case that non-neurological models are necessarily silent on neurological issues. For example, Nosofsky and Zaki (1998) accommodated differences in recognition and classification performance between amnesics and controls through changes in the overall confusability parameter of GCM (the *c* parameter). Nevertheless, it seems likely that due to the apparent functional specialization in

the brain, a model specified in neurological terms would be able to accommodate a broader range of independent variables than one that is not neurologically specified.

In addition, specification of a model in neurological terms opens up the possibility of employing dependent variables that would otherwise seem to be largely unavailable. For example, it is hard to see how ALCOVE, GCM, KRES, or the Simplicity model, could – in their current form – accommodate dependent variables from brain imaging. Non-neurological formal models need not be silent on the interpretation of imaging studies, but their contribution seems likely to be informal. For example, regarding the single vs. multiple systems debate in categorization, Gureckis et al. (in press) presented imaging evidence supporting a single system hypothesis of previous imaging data, which had been taken to imply a multi-system account of categorization (Reber et al., 2003).

A neuroscience specification of a formal model can also help leverage predictions in a broad range of fields. For example, the explicit system of COVIS critically involves the prefrontal cortex, which is generally considered to be a relatively late-maturing brain structure. This leads to the prediction that children should be particularly impaired relative to adults on categorization structures that COVIS indicates heavily involve the explicit system. There is some empirical support for this hypothesis (Minda, Desroches & Church, 2008).

Another possible reason for specifying formal models in terms of their instantiation in gross neuroanatomy is that neurological plausibility might, in itself, be a reasonable (if informal) method of model comparison. Many researchers from the mid 1980s onwards were convinced by the connectionist argument that, in the development of formal models of human behavior, it made sense to take into account what we know about the function of single neurons (McClelland & Rumelhart, 1985). It seems to us that, if one finds this connectionist line of argument convincing, one should also accept the same argument when it comes to gross neuroanatomy.

Counter-arguments and responses

In this section, we consider a number of possible counter-arguments to the suggestion that relative adequacy should be assessed by comparing well-defined models on the basis of the number and proportion of irreversible, ordinal, penetrable successes.

Feasibility

One concern is that, even when one ensures models are well-defined, the comparison of models on the number of ordinal irreversible successes represents a utopian ideal rather than a practical research strategy. For example, one might argue that the volume of research involved in achieving this form of comparison is larger than the scope of any single research article. One might also point to the fact that, even in a case of the evaluation of two models against a single contrast, this work often requires multiple experiments (controls for confounds, tests of generality, and so forth) and hence these narrow comparisons may represent a practical limit as to what can be achieved in any single publication.

Considered as a whole, the body of work required by the approach we advocate is clearly larger than a single article can accommodate. However, current practice of formal modelers in categorization indicates there are ways around this practical problem. Individual phenomena can be, and are, considered at length in primary empirical articles. Formal modeling may play an important role in these empirical articles – for example, if formal models differ in their ability to accommodate that single phenomenon, then the phenomenon under study may be of broader theoretical interest. Through these necessarily narrow empirical articles, and their associated follow-up work, one begins to establish a database that should permit researchers to identify *key phenomena*. These key phenomena then provide a canon of results against which broad model comparisons can be conducted.

Key phenomena are those that, through the rigors of close examination, are generally considered to be empirically robust and of likely theoretical importance. There is unlikely to ever be complete agreement about what constitutes a key phenomenon, but we are hopeful that it is possible to settle on a relatively large number of phenomena for which there is a broad consensus about their theoretical importance. This consensus might take the form of a list of phenomena that any new model of categorization should be expected to be able to accommodate. For example, the results of Shepard et al. (1961, see Figure 1), and of Medin and Schaffer (1978, see Figure 6) are strong candidates for such a list. We also anticipate that a much longer list is possible, but construction of such a list is beyond the scope of the current article.

Categorization modelers increasingly test their models against existing key phenomena, rather than attempt to derive new empirical data and test a model in the same study. For example, Love et al. (2004) evaluated the SUSTAIN model against 7 phenomena they considered to be key. Model parameters were held constant across the whole set of phenomena, so this paper represents 7 irreversible successes under the criteria proposed. Love et al. (2004) also formally considered the ALCOVE model against 4 of the 7 phenomena against which SUSTAIN had also been assessed (the remaining three phenomena were presumably considered to be outside ALCOVE's explanatory scope). Note that, in Love et al.'s evaluation of ALCOVE, model parameters were estimated independently for each of the four phenomena, so these do not constitute irreversible success as defined here. As ALCOVE was unable to capture one of these four phenomena (the results of Medin, Dewey & Murphy, 1983), whilst SUSTAIN was able to capture all four, the additional flexibility afforded ALCOVE in this case may have over-estimated its relative adequacy in the context of these four phenomena anyway. Of course, from the perspective of the current article, the central issue is that the comparison of two models against four phenomena is still a fairly

narrow comparison, relative to the number of "live" models and reliable phenomena within the domain of those models.

Another example that captures aspects of the approach we advocate is the work reported in Nosofsky (1992). In this publication, four models are formally compared against the results of more than thirty previously reported experiments. What is missing, from our perspective, is that (a) the comparison is limited to SSE adequacy, and (b) model parameters are determined independently for each comparison, and hence the consideration of multiple studies does not increase the number of successes, at least under the criteria proposed here. Given the existing work, a research article that combined the strengths of the approach taken by Love et al. (2004) and by Nosofsky (1992), which is the approach we advocate, seems like the next logical step. It may be that practical limits on the length of a single research paper mean that, as the number of reliable phenomena and candidate models increases, single papers will, of necessity, focus on either the establishment of new empirical phenomena, or on the relative adequacy of formal models across a broad range of well-established phenomena.

Another potential concern about the feasibility of the approach we advocate is the issue of whether it is practical to expect model parameters to remain constant across different experiments when those experiments involve, for example, different stimuli, and populations of participants. We have a number of responses to this concern.

First, our proposal represents a suggested goal for the formal modeling of categorization. The goal is a challenging one, and progress towards it is likely to be incremental rather than immediate. One might therefore anticipate the continued presence of arbitrarily variable parameters, at least to a limited extent, for some time to come.

Second, holding parameters constant across multiple experiments is clearly not impossible, as it has already been achieved in the work of Love et al. (2004). Perhaps one of

the reasons Love et al. (2004) were able to achieve this was their focus on the ordinal predictions of the model, rather than on precise quantitative fit. It may be that, at least for the current models, ordinal success is most that can reasonably be achieved in broad comparisons.

Third, our approach does not necessarily require that parameters are constant across experiments, simply that their values can be determined in a non-arbitrary way. For example, some stimulus dimensions are separable whilst others are relatively integral. The integrality of stimulus dimensions can be determined empirically in a number of ways (Garner, 1978), and this data can be used to determine in a non-arbitrary way whether a Euclidean or city-block metric is employed in a particular comparison of GCM to classification data.

The issue of differing populations of participants permits a couple of different responses. First, to the extent that different populations result in ordinally different patterns of results in the same experiments, this is presumably an important phenomenon in its own right. Accordingly. it should be possible, through further research, to identify the nature of the differences in such a way that they can be accommodated by models in a non-arbitrary manner (indeed, such investigations are essentially part of cross-cultural research; cf. Lopez et al., 1997) Alternatively, it would be possible to define the scope of models to include only particular populations. Second, if the focus of modeling is to account for the results of each individual, rather than just the group average (and strong arguments have been made for fitting models at an individual level, e.g. Ashby, Maddox & Lee, 1994; Smith, Murray & Minda, 1997; but see also Olsson et al., 2004), then it should be the case that an adequate model will have a fixed (or, at least, non-arbitrarily determined) set of parameter values for any given individual. This opens up the interesting possibility of the study of *super-participants* – in other words, the process of asking participants to complete a substantial battery of key phenomena across multiple sessions, and fitting one's model at the individual

participant level in terms of performance across that battery of results (assuming that these additional tests measure fairly stable characteristics).

As Olsson et al. (1994) have previously pointed out, one of the dangers of fitting individual participant, rather than average, data is that individual data can be noisier and hence more prone to overfitting. A sensible precaution is to compare models on their ability to fit both individual data and group averages; Hierarchical Bayesian Analysis provides one principled method for doing this (Shiffrin, Lee, Kim & Wagenmakers, 2008). Also, where possible, it seems sensible to collect sufficient data from each participant such that methods such as cross-validation can be employed at an individual participant level.

In summary, it would appear that there are several ways in which the dependence of models on arbitrarily variable parameters could be reduced and some researchers have adopted such approaches. Our point is that more effort should be invested in this direction, as it has the potential of greatly clarifying the issue of model comparisons.

Sufficiency of the current proposals

We have argued that progress in formal modeling would be improved by the comparison of well-defined penetrable models on the basis of the number of irreversible ordinal successes that can be attributed to them across a broad range of key phenomena. However, this is unlikely to be sufficient. For example, it is also important to seek clarity about what plausible results a model could not fit. We think that the practice of comparing models on the basis of the number of irreversible ordinal successes should, through the elimination of arbitrarily variable parameters, reduce model flexibility and hence decrease the likelihood that models derive their success through being able to fit any plausible result. However, it is doubtless important to verify that the models that emerge from this broad comparison process are able to make falsifiable predictions, and that the ordinal pattern produced by a model is observed in the experimental data (see Roberts & Pashler, 2000, for

further discussion of these important points). One potentially powerful method for examining whether models can be falsified is parameter space partitioning (Pitt, Myung, Montenegro & Pooley, 2008).

Differing goals

One defense of the continued existence of a large number of different formal models of categorization is that the different models reflect different research goals and hence the models are complementary rather than competing. For example, the modeling of reaction time distributions in the categorization of abstract items might be viewed as a different goal to the modeling of the effects of prior knowledge on classification probability. If the domains of these different models are well-defined, in the sense that the IVs and DVs to which they should and should not be applied is made explicit, then the models could indeed be considered to be complementary, and we have argued in this article (under the criterion of proportion of success) that the development of models with a relatively narrow explanatory scope might be one strategy in the process of developing adequate theories with a broad scope. Unfortunately, current formal models of categorization are typically ill-defined with respect to most DVs and IVs other than those to which they have been specifically applied. An explicit statement of the IVs and DVs to which a model is intended to apply would be a relatively straight forward way to assist progress in the formal modeling of categorization.

Other issues

Model selection statistics

In our discussion of the use of ordinal success in model comparison, we indicated that quantitative closeness may sometimes provide useful additional information. For the sake of brevity of expression, we focused this discussion on one of the simplest measures of quantitative fit – the sum of squared errors (SSE). There are a number of alternatives to SSE

that instantiate an Occam's razor approach to model selection. In other words, they aim to select the simplest model that provides an adequate explanation of the data. For example, the Akaike Information Criterion (AIC: Akaike, 1973) includes a term that penalizes the model on the basis of the number of free parameters it contains. The Bayesian Information Criterion (BIC: Shwarz, 1978) serves a conceptually similar function. AIC and BIC are increasingly employed in the quantitative evaluation of formal models of categorization and other phenomena (e.g. Ashby et al., 1998; Bergert & Nosofsky, 2007). Although neither AIC nor BIC take account of the fact that models with the same number of parameters can differ in flexibility, more recent Bayesian methods avoid this limitation (Myung & Pitt, 1997).

Overall, statistics that take into account model flexibility are to be preferred over SSE as measures of quantitative fit, When used in the comparison of models across a broad range of phenomena, they have to potential to redress the vast flexibility that comes from allowing a model to take a different parameter values for each data set it is applied to. However, model selection statistics do not obviate the need to consider ordinal success, nor the importance of making comparisons across a broad range of phenomena. Indeed, model selection statistics may be actively misleading when applied to a narrow data set. It is not difficult to devise situations where a model is more complex than it needs to be to accommodate a narrow data set even where, in the context of a broader set of data, the more complex model is the more adequate one.

Distinguishing representations from processes

In Figure 2, we provided a broad schema for the components of a formal model of categorization. Some of these components were described as representations and others as processes. Although this kind of componential analysis can make models easier to describe, in reality a model's behavior is often a product of subtle interactions between its components and it can be difficult to be sure which aspect of the model is responsible for its success (or

failure). For example, Rodrigues and Murre (2007) demonstrated that evidence taken to support the ATRIUM dual-system (rules and exemplars) model of categorization could also be accommodated by ALCOVE (a single-system exemplar model) if ALCOVE's attentional system was modified to allow exemplar-specific attention. Similarly, some results consistent with the cluster-based intermediate representations of the SUSTAIN model can also be accommodated by ALCOVE via exemplar-specific attention (Sakamoto, Matsuka & Love, 2004). The presence of such interactions provides another reason for comparing models against as broad a set of phenomena as possible. Where comparisons are narrow, there are likely to be a number of qualitatively different models that can accommodate the data under consideration. As the breadth of data increases, the number of models that are functionally equivalent with respect to that data seems likely to reduce. In the limiting case, two differently formulated models that capture all known reliable data are functionally equivalent and can be compared on aspects other than number of irreversible successes – for example, their penetrability.

Conclusion

Categorization research is notable in psychology for the extent to which it incorporates formal modeling of empirical phenomena, but the last 25 years have seen a proliferation of models. Many of these models are considered by their proponents to be "live" (i.e. not obviously superseded). These models cannot be considered to be complementary, because this would require a clear specification of their complementary domains – such a specification does not exist. They must therefore be considered to be alternative, competing accounts of a common set of empirical phenomena.

In principle, the formal specification of models in psychology should allow one to assess the relative adequacy of competing accounts, against well-informed empirical tests. Yet, empirical comparisons of the adequacy of formal models of categorization typically

focus on comparing two or three models on a very narrow data set (often a single experiment). Even setting aside model comparison, assessment of the adequacy of individual models is typically assessed against a few experiments at a time, with model parameters being derived independently on each occasion. This makes it difficult to ascertain whether these separately determined successes are compatible with each other – in the sense of being accountable within the same set of model parameters.

A necessary, but not sufficient, condition for progress in the formal modeling in psychology is to reach consensus on the criteria by which relative model adequacy should be assessed. Through a review of examples of good practice in the current categorization literature, we have concluded that progress is most likely to be achieved by comparing well-defined models on the basis of the number and proportion of irreversible, penetrable, ordinal successes that can be attributed to them across a broad range of key phenomena. Of the principles encapsulated in that statement, the irreversible nature of a model success is the most important – without that, formal modeling reduces to a more effortful version of informal verbal theorizing. Our definition of irreversible success is satisfied by counting successes with respect to evaluations employing global model parameters or principled accounts of how model parameters can be set in different situations. Probably the least important of the principles we propose is that of ordinal success. Other things being equal, better-than-ordinal precision is better than ordinal precision. But ordinal success seems an appropriate level to aim for at the current state of knowledge, and it avoids or reduces the risk and limitations inherent in relying on SSE for evaluating relative model adequacy.

Philosophically, the position we've taken in this article is most closely related to that of Lakatos (1978) rather than, for example, Popper (1934/1959) or Kuhn (1962). Lakatos (1978) argued that science differs from pseudoscience in that the former is able to make predictions about as-yet-unobserved phenomena. We argue that anyone wishing to claim that

the formal modeling of psychology could ever be a Lakatosian science should ensure they assess models in terms of irreversible successes. By so doing, models are more likely to be able to make clear predictions about novel phenomena and so restore confidence that the formal modeling of psychological phenomena is an endeavor worthy of the extensive effort and investment that it requires.

References

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

Anderson, J.R. (1991). The adaptive nature of human categorization. *Psychological Review, 98,*409-429.

Anderson, J.R., & Matessa, M. (1992). Explorations of an incremental, Bayesian algorithm for categorization. *Machine Learning, 9,*275-308.

Ashby, F.G., Alfonso-Reese, L.A., Turken, A.U., & Waldron, E.M. (1998). A neuropsychological theory of multiple systems in category learning. *Psychological Review, 105*,442-481.

Ashby, F.G., & Ell, S.W. (2001). The neurobiology of human category learning. *Trends in Cognitive Sciences, 5,*204-210.

Ashby, F.G., Ennis, J.M., & Spiering, B.J. (2007). A neurobiological theory of automaticity in perceptual categorization *Psychological Review, 114,*632-656.

Ashby, F.G. & Gott, R.E. (1988). Decision rules in the perception and categorization of multidimensional stimuli. *Journal of Experimental Psychology: Learning, Memory and Cognition, 14, 33-53.*

Ashby, F.G., & Maddox, W.T. (1993). Relations between prototype, exemplar and decision bound models of categorization. *Journal of Mathematical Psychology, 37,*372-400.

Ashby, F.G. & Maddox, W.T. (2005). Human category learning. *Annual Review of Psychology, 56,* 149-178.

Ashby, F.G., Maddox, W.T. & Lee, W.W. (1994). On the dangers of averaging across subjects when using multidimensional scaling or the similarity-choice model. *Psychological Science, 5,* 144-151.

Ashby, F.G., Paul, E.J. & Maddox, W.T. (2011). COVIS. In Pothos, E.M. & Wills, A.J. (Eds.). *Formal approaches in categorization.* Cambridge: Cambridge University Press.

Bergert, F.B., & Nosofsky, R.M. (2007). A response-time approach to comparing generalized rational and take-the-best models of decision making. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 33,* 107-129.

Boucher, L. & Dienes, Z. (2003). Two ways of learning associations. *Cognitive Science, 27,* 807-842.

Brooks, L. (1978). Nonanalytic concept formation and memory for instances. In E. Rosch & B.B. Lloyd (Eds.), *Cognition and Categorization*. Hillsdale, N.J.: Erlbaum.

Brooks, L.R., Squire-Graydon, R., & Wood, T.J. (2007). Diversion of attention in everyday concept learning: Identification in the service of use. *Memory and Cognition, 35*,1-14.

Bruner, J.S., Goodnow, J.J., & Austin, G.A. (1956). *A Study of Thinking*. New York: Wiley.

Chater, N. (1996). Reconciling Simplicity and Likelihood Principles in Perceptual Organization. *Psychological Review, 103*, 566-591.

Colreavy, E., & Lewandowsky, S. (2008). Strategy development and learning differences in supervised and unsupervised categorization. *Memory & Cognition, 36*,762-775.

DeCaro, M.S., Thomas, R.D., & Beilock, S.L. (2008). Individual differences in category learning: Sometimes less working memory capacity is better than more. *Cognition, 107*,284-294.

Delorme, A., Rousselet, G., Mace, M.J., & Fabre-Thorpe, M. (2000). Ultra-rapid categorization of natural scenes does not rely on colour cues: A study in monkeys and humans. *Vision Research, 40,*2187-2200.

Ennis, D.M. (1988). Confusable and discriminable stimuli: Comment on Nosofsky (1986) and Shepard (1986). *Journal Of Experimental Psychology: General, 117*,408-411.

Erickson, M.A., & Kruschke, J.K. (1998). Rules and exemplars in category learning. *Journal Of Experimental Psychology: General, 127*,107-140.

Fried, L.S., & Holyoak, K.J. (1984). Induction of Category Distributions: A Framework for classification learning. *Journal of Experimental Psychology, 10*,234-257.

Friston, K. (2010). The free-energy principle: a unified brain theory? *Nature Reviews Neuroscience, 11,* 127-138.

Garner, W. R. (1978). Aspects of a stimulus: Features, dimensions and configurations. In E. Rosch & B.B. Lloyd (Eds.), *Cognition & Categorisation*. Hillsdale, NJ: LEA.

Gluck, M.A. & Bower, G.H. (1988). From conditioning to category learning – an adaptive network model. *Journal of Experimental Psychology: General, 117,* 227-247.

Griffiths, O., & Mitchell, C.J. (2008). Selective attention in human associative learning and recognition memory. *Journal Of Experimental Psychology: General, 137*,626-648.

Gureckis, T.M., & Love, B.C. (2004). Common mechanisms in infant and adult category learning. *Infancy, 5*,173-198.

Gurcekis, T.M., James, T.W. & Nosofsky, R.M. (in press). Re-evaluating dissociations between implicit and explicit category learning: An event-related fMRI study. *Journal of Cognitive Neuroscience.*

Haslam, C., Wills, A.J., Haslam, S.A., Kay, J., Baron, R., & McNab, F. (2007). Does maintenance of colour categories rely on language? Evidence to the contrary from a case of semantic dementia. *Brain and Language, 103*,251-263.

Herrnstein, R. J., & Loveland, D. H. (1964). Complex visual concept in the pigeon. *Science, 146*,549-551.

Homa, D., Hout, M.C., Milliken, L. & Milliken, A.M. (2011). Bogus concerns about the false prototype enhancement effect, *Journal of Experimental Psychology: Learning, Memory and Cognition, 37,* 368-377.

Johansen, M.K. & Palmeri, T.J. (2002). Are there representational shifts during category learning? *Cognitive Psychology, 45,* 482-553.

Kemler, D.G. (1983). Exploring and reexploring issues of integrality, perceptual sensitivity and dimensional salience. *Journal of Experimental Child Psychology, 36*,365-379.

Kemler Nelson, D.G. (1984). The effects of intention on what concepts are acquired. *Journal of Verbal Learning and Verbal Behavior, 23*,734-759.

Knowlton, B.J., & Squire, L.R. (1993). The learning of categories: Parallel brain systems for item memory and category knowledge. *Science, 262,* 1747–1749.

Kruschke, J.K. (1992). ALCOVE: An exemplar-based connectionist model of category learning. *Psychological Review, 99,*22-44.

Kruschke, J.K. (2001). Toward a unified model of attention in associative learning. *Journal of Mathematical Psychology, 45*,812-863.

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

Lakatos, I. (1978). *The methodology of scientific research programmes: Philosophical papers volume 1*. Cambridge: Cambridge University Press.

Lamberts, K. (1995). Categorization under time pressure. *Journal Of Experimental Psychology: General, 124*,161-180.

Lamberts, K., & Freeman, R.P.J. (1999). Building object representations from parts: Tests of a stochastic sampling model. *Journal of Experimental Psychology: Human Perception and Performance, 25*,904-926.

Little, D.R., Nosofsky, R.M. & Denton, S.E. (2011). Response-time tests of logical-rule models of categorization. *Journal of Experimental Psychology: Learning, Memory and Cognition, 37,* 1-27.

López, A., Atran, S., Coley, J. D., Medin, D. L., & Smith, E. E. (1997). The tree of life: Universal and cultural features of folkbiological taxonomies and inductions. *Cognitive Psychology, 32,* 251-295.

Love, B.C., & Gureckis, T.M. (2007). Models in search of a brain. *Cognitive, Affective, & Behavioral Neuroscience, 7,*90-108.

Love, B.C., Medin, D.L., & Gureckis, T.M. (2004). SUSTAIN: A network model of category learning. *Psychological Review, 111,*309-332.

McClelland, J.L., & Rumelhart, D.E. (1985). Distributed memory and the representation of general and specific information. *Journal Of Experimental Psychology: General, 114*, 159-188.

McKinley, S.C., & Nosofsky, R.M. (1995). Investigations of exemplar and decision bound models in large, ill-defined category structures. *Journal of Experimental Psychology: Human Perception and Performance, 21,*128-148.

McKinley, S.C. , & Nosofsky, R.M. (1996). Selective attention and the formation of linear decision boundaries. *Journal of Experimental Psychology: Human Perception and Performance, 22*, 294-317.

McLaren, I.P.L., Leevers, H.L., & Mackintosh, N.J. (1994). Recognition, categorization and perceptual learning. In C. Umilta & M. Moscovitch (Eds.), *Attention & Performance* (Vol. XV). Cambridge, MA: MIT Press.

McLaren, I.P.L., & Mackintosh, N.J. (2000). An elemental model of associative learning: I. Latent inhibition and perceptual learning. *Animal Learning & Behavior, 28,*211-246.

Mackintosh, N.J. (1975). A theory of attention: Variations in the associability of stimuli with reinforcement. *Psychological Review, 82,*276-298.

"mass, n.[2]". (2011). In *Oxford English Dictionary.* Retrieved from http://www.oed.com:80/Entry/114666

Medin, D.L., & Atran, S. (2004). The native mind: Biological categorization and reasoning in development and across cultures. *Psychological Review, 111*,960-983.

Medin, D. L., Dewey, G. I., & Murphy, T. D. (1983). Relationships between item and category learning: Evidence that abstraction is not automatic. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 9,* 607–625

Medin, D.L., & Schaffer, M.M. (1978). Context theory of classification learning. *Psychological Review, 85*,207-238.

Milton, F., Longmore, C.A., & Wills, A.J. (2008). Processes of overall similarity sorting in free classification. *Journal of Experimental Psychology: Human Perception and Performance, 34*,676-692.

Minda, J.P., & Smith, J.D. (2001). Prototypes in category learning: The effects of category size, category structure, and stimulus complexity. *Journal of Experimental Psychology: Learning, Memory & Cognition, 27*,775-799.

Minda, J. P., & Smith, J. D. (2002). Comparing prototype-based and exemplar-based accounts of category learning and attentional allocation. *Journal of Experimental Psychology-Learning Memory and Cognition, 28*, 275-292.

Minda, J.P., Desroches, A.S., & Church, B.A. (2008). Learning Rule-Described and Non-Rule-Described Categories: A Comparison of Children and Adults. *Journal of Experimental Psychology-Learning Memory and Cognition, 34*,1518-1533.

Murphy, G.L. (2002). *The big book of concepts*. Cambridge, MA: MIT Press.

Murphy, G.L. (2011). The contribution (and drawbacks) of models to the study of concepts. In E.M. Pothos & A.J. Wills (Eds.), *Formal approaches in categorization*. Cambridge, UK: Cambridge University Press.

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

Newell, B., Dunn, J. C. & Kalish, M. (2010). The dimensionality of perceptual category learning: A state-trace analysis. *Memory & Cognition, 38,* 563-581.

Nosofsky, R.M. (1984). Choice, similarity, and the context theory of classification. *Journal of Experimental Psychology: Learning, Memory & Cognition, 10*,104-114.

Nosofsky, R.M. (1986). Attention, similarity and the identification-categorisation relationship. *Journal Of Experimental Psychology: General, 115*,39-57.

Nosofsky, R.M. (1987). Attention and learning processes in the identification and categorization of integral stimuli, *Journal of Experimental Psychology: Learning, Memory & Cognition, 13*, 87-108.

Nosofsky, R.M. (1988). Similarity, Frequency, and Category Representations. *Journal of Experimental Psychology: Learning Memory and Cognition, 14*,54-65.

Nosofsky, R.M. (1992). Exemplars, prototypes, and similarity rules. In A. Healy, S. Kosslyn, & R. Shiffrin (Eds.), *From Learning Theory to Connectionist Theory: Essays in Honor of William K. Estes, Volume 1* . Hillsdale, NJ: Erlbaum.

Nosofsky, R.M. (2000). Exemplar representation without generalization? Comment on Smith and Minda's (2000) "Thirty categorization results in search of a model". *Journal of Experimental Psychology: Learning Memory and Cognition, 26*,1735-1743.

Nosofsky, R.M., Gluck, M.A., Palmeri, T.J., McKinley, S.C., & Glauthier, P. (1994). Comparing models of rule-based classification learning: A replication and extension of Shepard, Hovland, and Jenkins (1961). *Memory & Cognition, 22*,352-369.

Nosofsky, R.M. & Kruschke, J.K. (1992). Investigations of an exemplar-based connectionist model of category learning. *The Psychology of Learning and Motivation, 28,* 207-250.

Nosofsky, R.M., & Kruschke, J.K. (2002). Single-system models and interference in category learning: Commentary on Waldron and Ashby (2001). *Psychonomic Bulletin & Review, 9*,169-174.

Nosofsky, R.M., Kruschke, J.K., & McKinley, S.C. (1992). Combining exemplar-based category representations and connectionist learning rules. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 18,* 211-233.

Nosofsky, R.M., & Palmeri, T.J. (1996). Learning to classify integral-dimension stimuli. *Psychonomic Bulletin & Review, 3*, 222-226.

Nosofsky, R.M., & Palmeri, T.J. (1997a). An exemplar-based random walk model of speeded classification. *Psychological Review, 104*,266-300.

Nosofsky, R.M., & Palmeri, T.J. (1997b). Comparing exemplar-retrieval and decision-bound models of speeded classification. *Perception & Psychophysics, 59,* 1027-1048.

Nosofsky, R.M., & Stanton, R.D. (2005). Speeded classification in a probabilistic category structure: Contrasting exemplar-retrieval, decision-boundary, and prototype models. *Journal of Experimental Psychology: Human Perception and Performance, 31,* 608-629.

Nosofsky, R.M., & Stanton, R.D. (2006). Speeded old-new recognition of multidimensional perceptual stimuli: Modeling performance at the individual participant and individual item levels. *Journal of Experimental Psychology: Human Perception and Performance, 32,* 314-334.

Nosofsky, R.M., & Zaki, S.R. (1998). Dissociations between categorization and recognition in amnesiacs and normals: An exemplar-based interpretation. *Psychological Science, 9*,247-255.

Nosofsky, R.M., & Zaki, S.R. (2002). Exemplar and prototype models revisited: Response strategies, selective attention, and stimulus generalization. *Journal of Experimental Psychology: Learning, Memory & Cognition, 28*,924-940.

Olsson, H., Wennerholm, P., & Lyxzen, U. (2004). Exemplars, prototypes, and the flexibility of classification models. *Journal of Experimental Psychology: Learning Memory and Cognition, 30*,936-941.

Palmeri, T.J., & Nosofsky, R.M. (2001). Central tendencies, extreme points, and prototype enhancement effects in ill-defined perceptual categorization. *Quarterly Journal of Experimental Psychology, 54A*,197-235.

Pearce, J.M. (1987). A model of stimulus generalization for Pavlovian conditioning. *Psychological Review, 94*,61-73.

Perry, C., Ziegler, J.C. & Zorzi, M. (2007). Nested incremental modeling in the development of computational theories: The CDP+ model of reading aloud. *Psychological Review, 114*, 273-315.

Pitt, M.A.., Kim, W., & Myung, J. (2003). Flexibility versus generalizability in model selection. *Psychonomic Bulletin & Review, 10*,29–44.

Pitt, M.A., Myung, J.I., Montenegro, M. & Pooley, J. (2008). Measuring model flexibility with parameter space partitioning: An introduction and application example. *Cognitive Science, 32*, 1285-1303.

Popper, K. (1934 / 1959). *Logik der Forschung*: Mohr Siebeck; 1959 English version: *The Logic of Scientific Discovery*: Routledge.

Pothos, E.M. (2005). The rules versus similarity distinction. *Behavioral and Brain Sciences, 28*,1-49.

Pothos, E.M., & Bailey, T.M. (2009). Predicting category intuitiveness with the rational model, the simplicity model, and the generalized context model. *Journal of Experimental Psychology: Learning, Memory & Cognition, 35*,1062-1080.

Pothos, E.M., & Chater, N. (2002). A simplicity principle in unsupervised human categorization. *Cognitive Science, 26*,303-343.

Pothos, E.M. & Close, J. (2008). One or two dimensions in spontaneous classification: A simplicity approach. *Cognition, 107,*581-602.

Pothos, E.M. & Wills, A.J. (Eds.) (2011). *Formal approaches in categorization.* Cambridge: Cambridge University Press.

Pothos, E.M., & Wood, R.L. (2009). Separate influences in learning: Evidence from artificial grammar learning with traumatic brain injury patients. *Brain Research, 1275,*67-72.

"quark, n.^{2}". (2011). In *Oxford English Dictionary*. Retrieved from http://www.oed.com:80/Entry/155969

Reber, P., Gitelman, D., Parrish, T. & Mesulam, M. (2003). Dissociating explicit and implicit category knowledge with fMRI. *Journal of Cognitive Neuroscience, 15,* 574-583.

Reed, S. K. (1972). Pattern recognition and categorisation. *Cognitive Psychology, 3,*382-407.

Rehder, B., & Murphy, G.L. (2003). A knowledge-resonance (KRES) model of category learning. *Psychonomic Bulletin & Review, 10,*759-784.

Rehder, B., & Hoffman, A.B. (2005). Eyetracking and selective attention in category learning. *Cognitive Psychology, 51,*1-41.

Rescorla, R.A., & Wagner, A.R. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In A.H. Black & W.F. Prokasy (Eds.), *Classical conditioning II: Current research* (pp.64-99). New York: Appleton-Century-Crofts.

Rissanen, J. (1978). Modeling by shortest data description. *Automatica, 14,*465–471.

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

Rodrigues, P.M. & Murre, J.M.J. (2007). Rule-plus-exception tasks: A problem for exemplar models? *Psychonomic Bulletin & Review, 14,* 640-646.

Rosch, E. & Mervis, C.B. (1975). Family Resemblances: Studies in the Internal Structure of Categories. *Cognitive Psychology, 7*,573-605.

Sakamoto, Y., Matsuka, T., & Love, B. C. (2004). Dimension-wide vs. exemplar-specific attention in category learning and recognition. In M. Lovett, C. Schunn, C. Lebiere, and P. Munro (Eds.), *Proceedings of the International Conference of Cognitive Modeling* (pp. 261-266). Mahwah, New Jersey: Lawrence Erlbaum.

Saksida, L.M. (1999). Effects of similarity and experience on discrimination learning: A nonassociative connectionist model of perceptual learning. *Journal of Experimental Psychology: Animal Behavior Processes, 25*,308-323.

Sanborn, A.N., Griffiths, T.L., & Navarro, D. (2006). A more rational model of categorization. In R. Sun & N. Miyake (Eds.), *Proceedings of the 28th Annual Conference of the Cognitive Science Society.*

Sanborn, A.N., Griffiths, T.L. & Navarro, D.J. (2010). Rational approximations to rational models: Alternative algorithms for category learning. *Psychological Review, 117*, 1144-1167.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics, 6,* 461–464.

Shepard, R.N. (1958). Stimulus and response generalization: Tests of a model relating generalization to distance in psychological space. *Journal of Experimental Psychology, 55*,509-523.

Shepard, R.N., Hovland, C.L., & Jenkins, H.M. (1961). Learning and memorization of classifications. *Psychological Monographs, 75*(13), Whole No. 517.

Shiffrin, R.M., Lee, M.D., Kim, W. & Wagenmakers, E. -J. (2008). A survey of model evaluation approaches with a tutorial on hierarchical Bayesian methods. *Cognitive Science, 32,* 1248-1284.

Shin, H.J., & Nosofsky, R.M. (1992). Similarity scaling studies of dot pattern classification and recognition. *Journal of Experimental Psychology: General, 121*, 278-304.

Smith, J.D. (2002). Exemplar theory's predicted typicality gradient can be tested and disconfirmed. *Psychological Science, 13*,437-442.

Smith, J.D., Beran, M.J., Crossley, M.J., Bloomer, J., & Ashby, F.G. (2010). Implicit and explicit category learning by macaques (Macaca mulatta) and humans (Homo sapiens). *Journal of Experimental Psychology: Animal Behavior Processes, 36*,54-65.

Smith, J.D., Chapman, W.P., & Redford, J. S. (2010). Stages of category learning in monkeys (Macaca mulatta) and humans (Homo sapiens). *Journal of Experimental Psychology: Animal Behavior Processes, 36*,39-53.

Smith, J.D., & Kemler Nelson, D.G. (1984). Overall similarity in adults' classification: The child in all of us. *Journal Of Experimental Psychology: General, 113*,137-159.

Smith, J.D., & Minda, J.P. (1998). Prototypes in the mist: The early epochs of category learning. *Journal of Experimental Psychology: Learning, Memory & Cognition, 24*,1411-1436.

Smith, J.D. & Minda, J.P. (2000). Thirty categorization results in search of a model. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 26*,3-27.

Smith, J.D., Murray, M.J. & Minda, J.P. (1997). Straight talk about linear separability. *Journal of Experimental Psychology: Learning, Memory and Cognition, 23*, 659-680.

Smith, L.B. (1989). A model of perceptual classification in children and adults. *Psychological Review, 96*,125-144.

Stanton, R.D., Nosofsky, R.M. & Zaki, S.R. (2002). Comparisons between exemplar-similarity and mixed-prototype models using a linearly separable category structure. *Memory & Cognition, 30*, 934-944.

Tharp, I. J. & Pickering, A. D. (2009). A note on DeCaro, Thomas, and Beilock (2008): Further data demonstrate complexities in the assessment of information–integration category learning. *Cognition, 111,*410-414.

Waldron, E.M., & Ashby, F.G. (2001). The effects of concurrent task interference on category learning: Evidence for multiple category learning systems. *Psychonomic Bulletin & Review, 8,*168-176.

Ward, T.B. (1983). Response tempo and separable-integral responding: Evidence for an integral-to-separable processing sequence in visual perception. *Journal of Experimental Psychology: Human Perception and Performance, 9,*103-112.

Widrow, B., & Hoff, M.E. (1960). *Adaptive switching circuits.* Paper presented at the IRE WESCON Convention.

Wills, A.J., Lavric, A., Croft, G.S., & Hodgson, T.L. (2007). Predictive learning, prediction errors, and attention: Evidence from event-related potentials and eye tracking. *Journal of Cognitive Neuroscience, 19,*843-854.

Wills, A.J., Lea, S.E.G., Leaver, L.A., Osthaus, B., Ryan, C.M.E., Suret, M.B., et al. (2009). A Comparative Analysis of the Categorization of Multidimensional Stimuli: I. Unidimensional Classification Does not Necessarily Imply Analytic Processing; Evidence From Pigeons (Columba livia), Squirrels (Scurius carolinensis), and Humans (Homo sapiens). *Journal of Comparative Psychology, 123,*391-405.

Wills, A.J., Longmore, C.A., & Milton, F. (2011). Does deliberative thought employ a smaller subset of the available information than non-deliberative thought? Some evidence to the contrary from studies of free classification. Submitted to *Journal of Experimental Psychology: Learning, Memory & Cognition*

Wills, A.J., & McLaren, I.P.L. (1997). Generalization in human category learning: A connectionist explanation of differences in gradient after discriminative and non-

discriminative training. *The Quarterly Journal of Experimental Psychology, 50A*,607-630.

Wills, A.J., & McLaren, I.P.L. (1998). Perceptual learning and free classification. *Quarterly Journal of Experimental Psychology, 51B*,235-270.

Wills, A.J., Reimers, S., Stewart, N., Suret, M., & McLaren, I.P.L. (2000). Tests of the ratio rule in categorization. *Quarterly Journal of Experimental Psychology, 53A*,983-1011.

Wills, A.J., Suret, M., & McLaren, I.P.L. (2004). The role of category structure in determining the effects of stimulus preexposure on categorization accuracy. *Quarterly Journal of Experimental Psychology, 57B*,79-88.

Zeithamova, D., & Maddox, W.T. (2006). Dual-task interference in perceptual category learning. *Memory and Cognition, 34*,387-398.

Footnotes

1. 2604 citations for 19 GCM papers published by Nosofsky. ISI Web-of-Science, 30-Nov-2009.
2. The design of the procedural-learning system in COVIS has changed over time; we concentrate here on the Striatal Pattern Classifier version (Ashby, Ennis & Spiering, 2007).
3. The extent to which the data sets considered by Smith and Minda were coherent and statistically independent is a matter of debate (Nosofsky, 2000).
4. It is important to stress that this approach to modeling the result of Wills et al. (2004) is not something that has been advocated by proponents of GCM – it is provided merely as hypothetical but particularly clear illustration of the problem of arbitrarily variable parameters.
5. Another possible solution would be to extend or modify the GCM to include a well-defined process by which *c* changes with exposure; see McLaren & Mackintosh, 2000 for one example of a formal model of exposure effects.
6. 11 articles with topic categorization in the 2009 issues of JEP:LMC. 2686 articles with topic categorization; search limited to psychology and neuroscience (full search criteria available on request). ISI Web-of-Science, 28-Jan-2010.

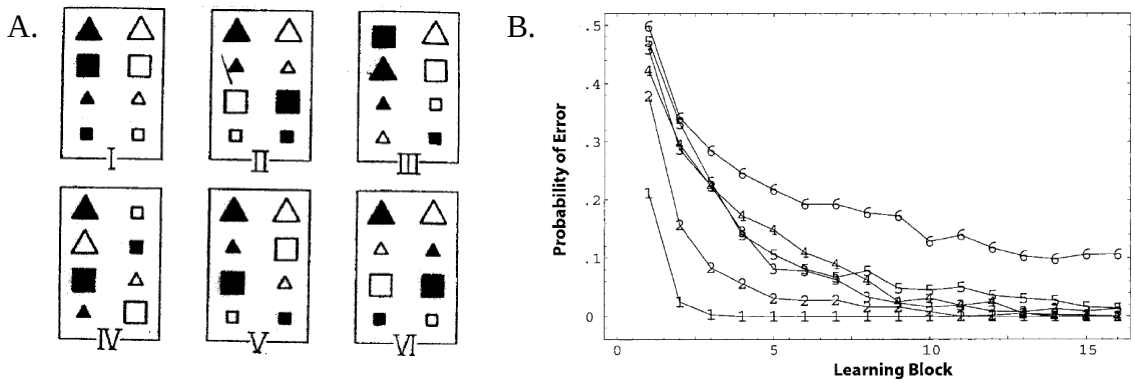

Figure 1. (A) Six different ways to classify eight stimuli into two different groups; the stimuli take one of two values on each of three dimensions (color, size, shape). (B) Mean errors in learning these category structures from feedback, as a function of amount of feedback (learning block). Figure 1A reproduced from Shepard, Hovland & Jenkins (1961), Figure 1B reproduced from Love, Medin and Gureckis (2004).

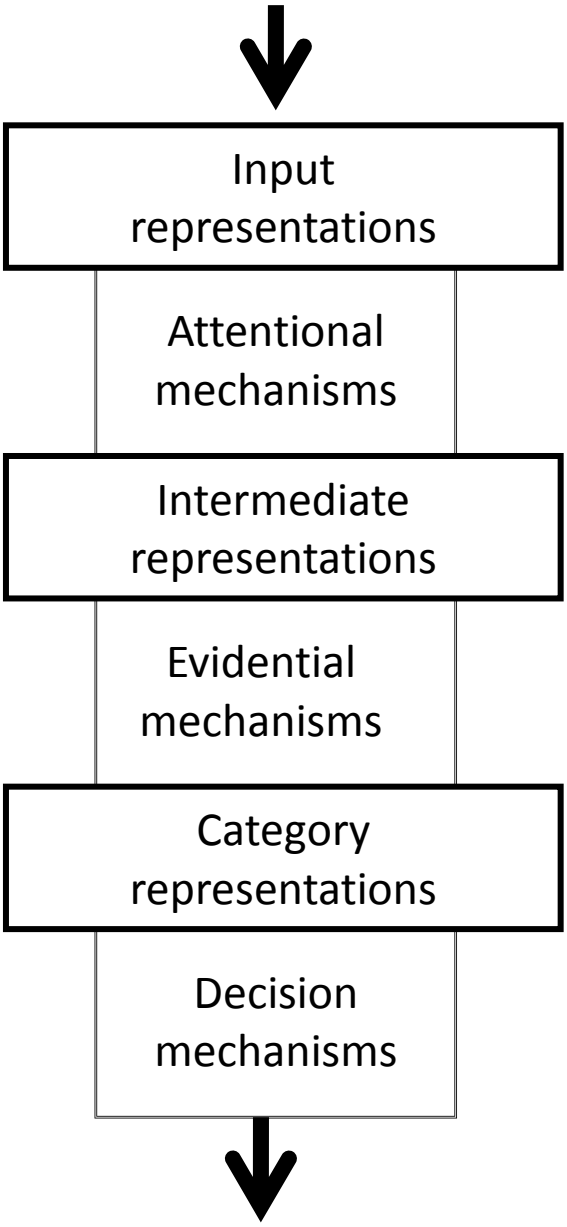

Figure 2. Schematic illustration of the typical representations and mechanisms of a formal model of categorization.

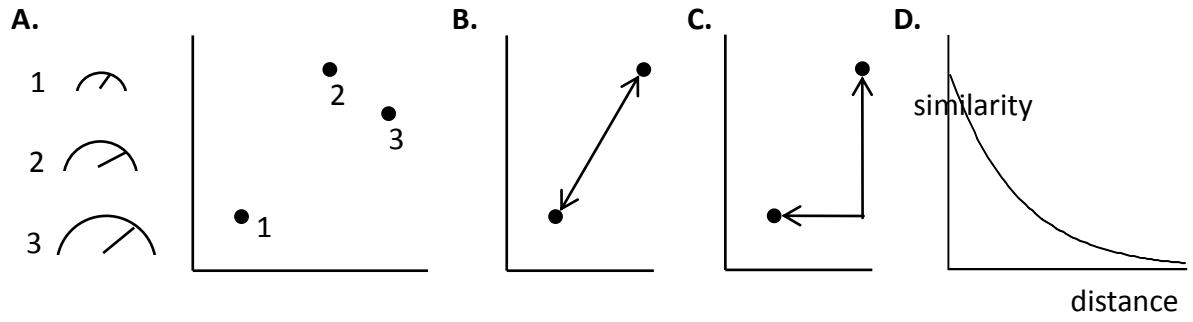

Figure 3. **A.** Representing the similarity structure of stimuli 1, 2 and 3 in a 2-dimensional geometric space; in this example the dimensions of this space are readily interpretable as size and angle**. B.** Euclidean distance ($distance^2 = x^2 + y^2$). **C.** City-block distance (distance = x + y). **D.** An exponential decay relationship between similarity and distance in psychological space.

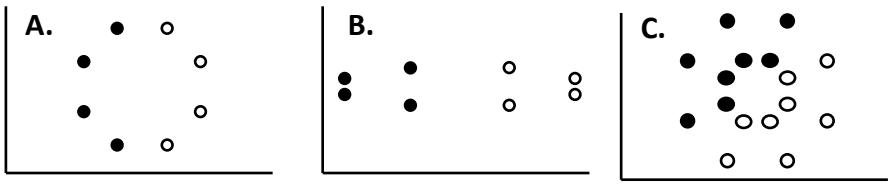

Figure 4. **A**. Geometric representation of two categories, each of four stimuli (category membership denoted by colour of dot). **B.** Stretching along the x-axis and compression along the y-axis, thereby increasing within-category similarity and decreasing between-category similarity. **C.** Overall expansion of psychological similarity space.

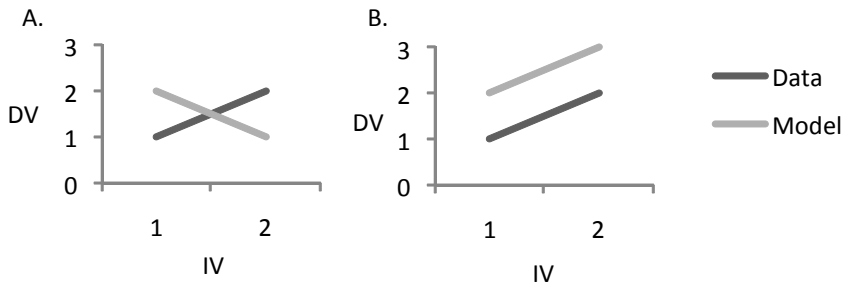

Figure 5. **A**. An ordinally incorrect model; sum of squared error (SSE) = 2. **B.** An ordinally correct model; SSE =2.

	C	F	S	N
A1	1	1	1	0
A2	1	0	1	0
A3	1	0	1	1
A4	1	1	0	1
A5	0	1	1	1

	C	F	S	N
B1	1	1	0	0
B2	0	1	1	0
B3	0	0	0	1
B4	0	0	0	0

Figure 6. The "5-4" category structure employed by Medin and Schaffer (1978). Category A has 5 items, category B has 4 items. C, F, S and N refer to the dimensions of color, form, size, and number. Each dimension takes one of two values, indicated by 1 or 0 in the above table.