The Right Tool for the Job: Information-Processing Analysis in Categorization

Kevin A. Gluck (kevin.gluck@williams.af.mil)
Air Force Research Laboratory, 6030 S. Kent St., Mesa, AZ 85212 USA

James J. Staszewski (jjs@andrew.cmu.edu)
Department of Psychology, Carnegie Mellon University, Pittsburgh, PA 15213 USA

Howard Richman (howard@pahomeschoolers.com)
Department of Psychology, Carnegie Mellon University, Pittsburgh, PA 15213 USA

Herbert A. Simon (hs18@andrew.cmu.edu)
Department of Psychology, Carnegie Mellon University, Pittsburgh, PA 15213 USA

Polly Delahanty (pd2w@andrew.cmu.edu)
Department of Psychology, Carnegie Mellon University, Pittsburgh, PA 15213 USA

Abstract

Smith and Minda (2000) showed that mathematical approximations of several popular categorization theories could be fit equally well to the average "percentage of 'A' responses" in their meta-analysis of studies that used the 5-4 category structure. They conclude that the 5-4 category structure is not a useful paradigm for explaining categorization in terms of cognitive processes. We disagree with their conclusion, and contend instead that the problem lies with the data collection and analysis methods typically used to study categorization (in this and other category structures). To support this claim, we describe a recently completed study in which we collected and used a variety of converging data to reveal the details of participants' cognitive processes in a 5-4 category structure task.

The Smith and Minda (2000) Meta-Analysis

Recently, Smith and Minda (2000) reanalyzed 29 data sets (each set a particular condition in an experiment) collected from the experimental literature on categorization that used the 5-4 category structure.[1] Eight of the sets employed the stimuli called Brunswik faces. Others used yearbook photos (4 sets), geometric shapes (11 sets), verbal descriptions (3 sets), and rocketships (3 sets).

As shown in Table 1, each stimulus in this category structure has 4 binary features, whose combination creates 16 (2^4) different stimuli. The 5-4 structure splits this set into two linearly-separable groups.

In the acquisition phase of a typical category learning study, participants first learn to classify 9 of the 16 stimuli, 5 as A and 4 as B, as shown in the table. Each trial presents the 9 learning items, one at a time, in a random sequence, and the order changes from trial to trial. Participants classify each as "A" or "B" and the correct assignment for each stimulus is given as feedback. Typically, learning proceeds until a participant classifies all 9 stimuli correctly in a single trial. In the transfer test that follows, all 16 stimuli are presented to the participants and they classify each, now without feedback.

Table 1: The 5-4 category structure.

Stimulus (M&S, 1981)	Stimulus (S&M, 2000)	Feature F1	F2	F3	F4
Category A					
4A	A1	1	1	1	0
7A	A2	1	0	1	0
15A	A3	1	0	1	1
13A	A4	1	1	0	1
5A	A5	0	1	1	1
Category B					
12B	B6	1	1	0	0
2B	B7	0	1	1	0
14B	B8	0	0	0	1
10B	B9	0	0	0	0
Transfer					
1A	T10	1	0	0	1
3B	T11	1	0	0	0
6A	T12	1	1	1	1
8B	T13	0	0	1	0
9A	T14	0	1	0	1
11A	T15	0	0	1	1
16B	T16	0	1	0	0

Note. M&S = Medin & Smith; S&M = Smith & Minda. The feature structure for Medin & Smith's stimulus 4 is identical to that in Smith & Minda's stimulus A1, and so on.

Fitting the Data

For purposes of their meta-analysis, Smith and Minda (2000) used data from the transfer trial in each of these 29

[1] They speak of 30 sets, but two of their sets are obviously a duplication from an experiment by Medin and Smith (1981), for all 16 data points are identical for both sets.

data sets. Specifically, for each of the 16 stimuli, they computed the percentage of participants in each study who classified a stimulus as an 'A' stimulus. They then averaged these percentages over all 29 data sets to provide a global average, containing 16 data points, one for each stimulus. Thus, each data point represents the average percentage of participants (in those 29 studies) who categorized a particular stimulus as an 'A' stimulus. These data are displayed in Figure 1.

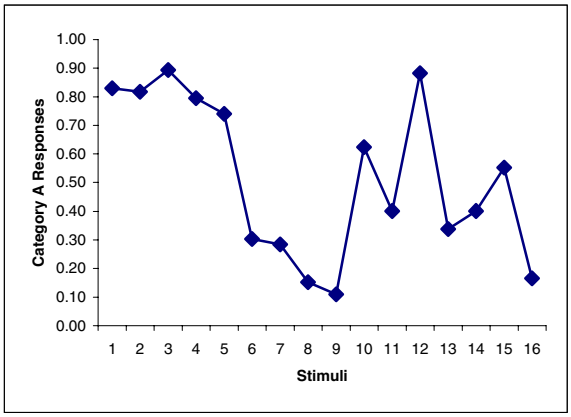

Figure 1. Average percentage of category 'A' responses from the Smith and Minda (2000) meta-analysis.

Smith and Minda (2000) next built a set of eight mathematical models, one for each of the theories they evaluated, and fitted each to the data for the 16 stimuli, adjusting the available parameters (4-6 free parameters per model) separately for best fit to each set of data. Then they averaged the set of 29 predicted performance profiles, to provide a set of 16 data points for comparison with the aggregated average of the actual data over all 29 studies. As Smith and Minda give a very clear and complete account of the functions they used and the fitting procedure, we need not repeat that information here. Table 2, adapted from Smith and Minda's Table 2, summarizes the fits to the data and the number of free parameters available to each model.

Table 2: Measures of Fit for Mathematical Models in Smith and Minda (2000)

Model	AAD	PVA	FP's
Additive prototype	0.091	0.838	4
Multiplicative prototype	0.069	0.890	5
Context	0.047	0.941	5
Additive exemplar	0.144	0.664	4
Fixed low sensitivity	0.149	0.637	5
Gamma	0.045	0.944	6
Twin sensitivity	0.043	0.946	6
Mixture	0.046	0.944	5

Note. AAD = Average Absolute Deviation; PVA = Percentage of Variance Accounted for (R^2); FP's = free parameters.

Models of the Experimental Materials

In the 5-4 category structure, individual features of value 1 (assignment of 1 and 0 is arbitrary in a particular experiment) dominate the five A stimuli and features of value 0 dominate the B stimuli: A stimuli average 2.8 1-features; B stimuli average 1.25. It is reasonable that participants would learn that 1-values indicate probable A membership, and 0 values, probable B membership. We refer to this characteristic of stimuli in the 5-4 category structure as "A-proneness."

Suppose we assign to each stimulus a value corresponding to its "A-proneness." For example, we could assign a stimulus 4 points for a 1 in F1, 3 points for a 1 in F2, 4 points for a 1 in F3, and 3 points for a 1 in F4, corresponding to the frequency with which these feature values are associated with category A in the learning set. Then take the sum of these values as the measure of A-proneness for a particular stimulus. By this method, stimulus 4A gets a score of 11, and stimulus 12B gets a score of 7.

We can fit this measure of A-proneness to the Smith and Minda (2000) data (percentage of category A responses for each stimulus) using a linear regression. The result is a 2-parameter model ($y = .034x + .069$) that predicts the percentage of category 'A' responses for all 16 stimuli with an R^2 of .81 and an AAD of .098. This simple linear regression, with only two free parameters, and derived strictly from the feature structure of the stimuli, accounts for the lion's share of the performance variability.

This analysis does not instill confidence that the more elaborate models have much to say about the actual psychological processes of the participants in these experiments. The only thing any of these models (including our strawman model) tell us about participants in these studies is that they are capable of tuning their behavior to the feature structure of the stimuli. These are more models of the experimental materials than they are models of psychological processes.

Indeed, Smith and Minda (2000) observe that the best of the mathematical models all produced such good fits to the data that it was impossible to choose between the very different process models motivating them:

> ". . . the underlying representation and process remains undetermined and unknown. Therefore, one sees that the [29] 5-4 data sets, when described by formal models, are silent on the matter of whether categories are represented in a way that is based on prototypes or in a way that is based on exemplars." (p. 17)

Smith and Minda (2000) conclude from this that the 5-4 category structure is too limited in its properties for general conclusions to be drawn from it about the processes that people use to learn new categories.

We do not agree that the 5-4 category structure is inherently too limited to reveal the underlying cognitive processes. We propose instead that the methods dominating this area of research are too limited. You've got to have the right tool for the job. In this case, the job is to uncover the processes underlying category learning and categorization

performance. We claim that the right tool (methodology) for this job is fine-grained information-processing analyses, using a variety of converging measures. We have found that detailed analysis of the trial-by-trial behaviors of individual participants reveals rich complexity in their categorization processes. In the next section, we describe the completed study, our approach to data analysis, and the lessons we have learned about the complexity and variability of categorization processes in the 5-4 paradigm.

An Information-Processing Analysis

The study we completed involved a partial replication of Medin and Smith (1981), whose category learning study implemented the 5-4 category structure using Brunswik faces. The four binary features of Brunswik faces are Eye Height (EH – High and Low), Eye Spacing (ES – Wide and Narrow), Nose Length (NL – Long and Short), and Mouth Height (MH – High and Low). These features – EH, ES, NL, and MH – correspond to features F1-F4, respectively, in Table 1. Like Medin and Smith, we used three instruction conditions (Standard, Prototype, and Rule-X), a learning phase, and a transfer phase. After that, participants were presented each of the 16 faces and its associated category, one at a time, and they rated the extent to which the face was typical for that category on a 1-to-9 scale. At the end, participants gave retrospective reports describing the processes they used to categorize the stimuli. Thirty-six Carnegie Mellon University undergraduates participated in this study. Half gave concurrent verbal protocols during the entire learning phase. Our analyses focus on these 18 participants.

A Variety of Measures

The data we have focused on in our process analyses include (1) errors, (2) concurrent verbal protocols, (3) typicality ratings, and (4) retrospective reports. Data analysis were not limited to measuring the frequency with which participants choose A or B responses to the 16 stimuli during a transfer trial. Instead, we relied on analysis of the detailed behavior of participants while they were performing both the learning and the transfer task: data that revealed a great deal about the processes they were using.

In performing these analyses, we have been guided by the idiographic data analysis methods typified by Newell and Simon (1972) and by a general theory of perception and memory, EPAM, that has previously been applied to aggregate data on the 5-4 task (Gobet, Richman, Staszewski, & Simon, 1997). EPAM is a computer program that uses a discrimination net architecture to simulate the participants' behavior in responding to each stimulus. In fact, EPAM was used to simulate the aggregate data from Medin and Smith (1981) that comprises three of the 29 Smith and Minda (2000) data sets.

Following are descriptions of our data analysis procedures, accompanied by illustrations of how analysis at this level of detail can reveal participants' categorization processes.

Errors. Participants may use from one to four features to classify a face, and they exercise most of these options at one time or another. Table 3 shows the likely errors that arise (out of ambiguities) when the nine faces used in the learning trials are categorized only on the basis of particular features, or particular pairs, or triplets of features.

The four rows and the first four columns of the table name the features. Each of the cells in the first four columns corresponds to a classification of the faces on the corresponding pair of features. For convenience of reference, we have designated the cells of the table corresponding to particular combinations of feature tests with letters from M through Z.

For example cell V, which is at the intersection of row EH and column NL, shows on the first line that 5A and 2B cannot be distinguished using only these two features, for both faces have identical eye heights and nose lengths (EH=0; NL=1). Similarly (second line), 13A and 12B are identical on EH (1) and NL (0), so a discrimination process that relied only on those two features would not be able to discriminate stimuli 13A and 12B. The other five faces form two classes: A's with high eyes and long noses, and B's with low eyes and short noses. So, for this particular pair of tests, four faces are ambiguous or "hard," and likely to be misclassified during learning. If, instead of EH:NL, the features attended to were nose length and mouth height (NL:MH), then 4A, 7A and 2B would fall in a single class, as would 13A and 14B, and these five would be the hard faces in this case. Thus, when participants use a particular pair of features to classify faces, they will make the greatest number of errors in classifying the faces that are hard for that pair.

The column of Table 11 marked "EXCEPT" indicates which faces would be error-prone if the three features *except* the one labeling the corresponding row were tested (i.e., a discrimination net using the three remaining features); the column marked "SOLE" indicates which faces would be error-prone if *only* the feature on that row were tested. The EXCEPT column shows that all nine faces can be categorized perfectly without the use of ES, but the three other features, EH:NL:MH, must all be used. Notice that each of the three triads of features that includes ES produces a different set of hard faces, as does each net using only a particular single feature.

By assessing which were the hardest faces during the learning phase, we identified the dominant discrimination strategy for each participant. Participants in the Prototype instruction condition showed the most between-subject variability in process, with strategies V, W, Y, P, and R inferred from their errors in the learning phase. Standard participants also showed considerable variability, with evidence of strategies V, W, Y, and Z in their data. The Rule-X participants, who were told explicitly to attend to nose length, used strategies R and V.

Table 3: Error patterns predicted by feature selection in the 5-4 categorization paradigm

	EH (F1)	ES (F2)	NL (F3)	MH (F4)	EXCEPT	SOLE
EH (F1)		5A,2B 4A,13A,12B **U**	5A, 2B 13A,12B **V**	5A,14B 4A,7A,12B **W**	4A,2B **M**	5A, 12B **N**
ES (F2)			4A,5A,2B 13A,12B **X**	15A,14B 7A,10B 4A,2B,12B **Y**	**O**	7A,15A 2B,12B **P**
NL (F3)				4A,7A,2B 13A,14B **Z**	4A,12B **Q**	13A,2B **R**
MH (F4)					13A,12B 5A,2B **S**	4A,7A 14B **T**

Note. Stimuli listed in each cell (e.g., 5A, 12B) are those for which errors are expected if the participant is attending to that conjunction or disjunction of features. Bold code letters (e.g., **U**, **V**, **W**) are used in the text as an economical means of referring to specific categorization strategies, as indicated by increased attention to specific features. F1-F4 = Features 1-4 in Table 1. EH = Eye Height; ES = Eye Spacing; NL = Nose Length; MH = Mouth Height. EXCEPT = attention to all features except the feature in that row. SOLE = attention to only the feature in that row.

Verbal Protocols. We assume that the features used in discriminating and categorizing the faces are verbalizable. The claim is not that participants will verbalize every feature to which they attend, or even that discrimination is always a verbal process; the claim is that the process of encoding features can create a verbalizable representation, and that patterns of verbalization of features are correlated with patterns of attention to the stimuli.

The Brunswik faces are easy to distinguish visually, and to describe verbally, using either the “official” features (eye height, eye spacing, nose length, mouth height) mentioned in the experimental instructions, or other descriptors that may be already familiar to individual participants(e.g., "long face", "small distance between nose and mouth," "wide face," or even "monkey-like"). The official descriptors, rather than idiosyncratic ones, are by far the more frequent in the protocols.

Participants' protocols mainly reported values of features of the face they were currently categorizing, sometimes supplemented with a reason for assigning the face to a particular class, and sometimes with a comparison with a previous face. The following (each preceded by identifier of participant and experimental condition) are examples of verbal responses to stimuli that described features in the language of the instructions:

MS (prototype). "High eyes; short nose; low mouth. Let's go with B, because the last one had high eyes and low mouth."

ML (prototype). "I'll say this is A because of the nose length and the eye height and the separation between the eyes and the mouth."

Rather more austere and more typical are:

MK (standard). "close and high eyes, small nose, middle mouth." *(Chooses A.)*

RB (prototype). "The eyes are low and the nose is big." (*Chooses B*.)

The discrimination processes of participants who use idiosyncratic descriptive terms are harder to identify, but the descriptors they actually used were generally related to the "official" ones in simple ways. For example, "long" faces were faces with high eyes, and sometimes also with low mouths. Faces with "eyes close to the nose" were faces with low, close eyes. The meaning, in terms of features, of these non-standard descriptors can usually be determined by checking the characteristics of the faces to which participants applied them.

Typicality Ratings. Following the transfer phase of the experiment, participants were shown each face along with its correct category. Their task was to rate the extent to which each face was typical for its category. The verbal protocol participants also provided explanations for their ratings. These proved to be informative as additional converging evidence regarding how participants were discriminating the stimuli. Following are several examples from the typicality rating explanations:

JIS (standard). “This one is pretty typical of A because the eyes are way up high and spread out in this one.”

RB (prototype). “That one, I think, is typical because the eyes are high and far apart, and the nose is little.”

MB (rule-x). “Typical. Short nose.”

In addition to lists of features as justification for the ratings, participants occasionally referred to Gestalt characteristics of the faces, using terms like "long" or "wide". For instance, "This one doesn't look like a B face. The eyes are high, and it looks like a kind of long face." The majority of explanations, however, were feature lists that revealed the various ways participants used combinations of feature values to categorize the stimuli.

Retrospective Reports. After the typicality ratings, the experimenter asked each participant, "On what basis were you making your classifications?" Following are two example responses:

RB (prototype). "Most of the type A had high eyes, and it didn't matter where the nose is or the mouth. And most of the B's had eyes in the middle, but there was a type B that had really high eyes. And then there was a type A that had eyes in the middle with a little nose and a long mouth."

MK (standard). "Basically, small nose was A, big nose was B. Basically, except small nose if the mouth was low, I looked at the eyes, and if the eyes were low, then it was B. If it was a big nose with little mouth and high up, then I checked the eyes, and if the eyes were high, then the face was A."

Note that both RB's and MK's retrospective reports are consistent with their concurrent verbalizations from the learning trials. It is converging evidence of this sort that increases our confidence in conclusions regarding participants' categorization processes.

Comparison of VP and NVP Errors

An assumption that is required in drawing generalizations from the verbal protocol participants is that the requirement to give protocols does not itself have a direct impact on categorization processes in this task. It would be reassuring if the performance of the verbal protocol (VP) participants and the non-verbal protocol (NVP) participants were in fact similar.

Following the logic in Table 3, to the extent participants in the VP and NVP conditions found similar faces difficult during the learning phase, there is evidence for similar categorization strategies across those conditions.

In both conditions, the four most difficult stimuli are 2, 5, 12, and 13. These are difficult stimuli because they are exceptions on features that are highly predictive of category membership. Table 1 shows that stimuli 2 and 13 are exceptions on Nose Length, while 5 and 12 are exceptions on Eye Height. Additionally, pairs of those stimuli are confusable if one ignores Mouth Height. That is, 2 and 5 are identical except for mouth height, as are 12 and 13. The fact that these four stimuli are always the most difficult, suggests that most participants, regardless of verbal protocol condition, found it difficult to learn the exceptions.

Looking at error rates across all of the stimuli, we find that VP and NVP error rates correlate $r = .82$, indicating a high degree of similarity in the error patterns between the two conditions. In terms of overall error rate, VP participants tended to make more errors (Mean = 50.2) than NVP participants (Mean = 39.7), although an ANOVA reveals that this difference is not significant: $F(1, 36) = 1.219$, $p = .277$).

Summary of Findings

Due to space limitations, and because of the massiveness and complexity of the body of data we are examining, we must briefly summarize our most important findings. Additionally, we feel that a general description of the results will be more useful, in terms of distinguishing the information processing approach from the more typical, aggregate-level, nomothetic approach, than the specific frequencies and percentages in our findings would be. Therefore, we will finish the paper with an account of our main findings.

One lesson that emerges from these rich data is that the task structure itself was a major determinant of the outcomes we measured. The influence of task structure on performance is apparent in Figure 2, and we found similar effects in our data. Because nearly all of our participants achieved the learning criterion, we infer that they discovered the implicit task structure.

A second finding is that most of the participants, regardless of their instructional condition, interpreted the task as one of forming rules that could be used to assign faces to a category. This generally took the form of learning what *features* were associated with the A or B categories, then using this knowledge to classify *faces* by means of their features, rather than defining prototypic faces or cycling through comparisons with previous exemplars. Feature-based rule following is apparent in the verbal protocols presented earlier.

Regarding rule-based behaviors, some (but not all) participants discovered that it was also useful to recognize certain individual faces and associate their categories directly with them. These were almost always the "hard" faces that did not fit the simple rules they used to discriminate the others. This was particularly apparent in statements like, "Ah, this is the one that tricks me." This phenomenon is consistent with the model of Nosofsky, Palmeri, and McKinley (1994), but the inconsistent appearance of this phenomenon in our data also suggests that model's limitations.

Another finding is that the numerous differences in the behaviors of different participants could be traced in large measure to different strategies of attention, and different strategies for retaining and combining information about features and combinations of features upon which rules of choice could be built. Strategies were effective to the extent that they made only modest demands on memory, including demands on short-term memory and demands for transferring information to long-term memory and retaining it for use in building up the structure of decision rules.

Perhaps the most consistent phenomenon we observed in the data was the high degree of within-participant variability in process. Examination of participants' stimulus-by-stimulus category responses and corresponding

verbalizations reveals that the dominant rule-based processes are sprinkled with instances of comparing the current stimulus to the immediately preceding stimulus, and also increasing evidence of recognition-based processes (especially for the hard faces) with experience. Even within the rule-based processes, there was a good deal of variability, as participants had varying degrees of success with the feature-based rules they generated and tested.

Conclusion

Sciences are concerned with discovering and testing laws that describe the invariant features of their domains. Invariance is a complex concept. Even the gravitational constant is not an invariant once one strays from the Earth or ascends a mountain. So science has laws, like the law of gravitational attraction, but it has parameters and variables that specify the workings of each law as a function of various circumstances.

Matters become especially complex when we consider laws of biology, with its immense variety of living forms, and still more complex when we consider the laws of psychology, which seeks the regularities in the behavior of an organism that has enormous capabilities for adaptation and learning. Not only will the behavior vary with the innumerable features of the environment in which the person performs the task, but even leaving genetic differences aside, it will vary as a function of each individual's previous history of experience and instruction.

The experimental data we have analyzed here illustrate a number of such complexities. No unitary set of laws, taken by itself, governs the precise way in which a set of people go about solving a simple categorization task, not even if all of them are drawn from the same university population. Covering variation by averaging conceals it but does not banish it or explain it. Siegler (1987) made the same point about the "Perils of Averaging," but in the context of analyzing children's arithmetic. With this paper we illustrate the value of a fine-grained, multivariate approach for advancing our understanding of categorization and category learning in terms of their underlying cognitive processes.

It was almost a half century ago that Bruner, Goodnow, and Austin (1956) published their seminal work on categorization, *A Study of Thinking*. In the introduction to that book, they wrote:

> ". . . we have come gradually to the conclusion that what is most needed in the analysis of categorizing phenomena . . . is an adequate analytic description of the actual behavior that goes on when a person learns how to use defining cues as a basis for grouping the events of his environment." (p. 23)

To understand participants' actual behaviors during categorization and category learning processes, not only did we need to analyze the behavior of individual participants, but the data obtained from each had to be of a grain size fine enough to capture some detail of ongoing learning processes. We examined errors and concurrent verbalizations stimulus-by-stimulus, then looked for converging evidence in participants' typicality ratings and their retrospective reports. The requirement that conclusions about participants' processes should be based on the convergence of multiple measures provided strong tests of the validity of our findings. In the end, our approach yielded a rich, descriptive understanding of the underlying representations and processes employed by participants in our study.

The challenge to explain the phenomena observed remains. Because of the variety of measures used in our analyses, and the variation among them in granularity, we advocate the development of simulation models. We submit that detailed, multivariate information-processing analyses and simulation modeling are tools that are well-suited for the job of advancing understanding of category learning and categorization. The data help us come to a better understanding of actual cognitive processes in category learning, and simulation models allow for the possibility of accounting for the enormous variability in process within and between participants.

Acknowledgements

The authors would like to thank Damian Bierman for assisting in the early stages of this research project. The opinions expressed in this paper are those of the authors and do not reflect the official policy or position of the United States Air Force, the U.S. Department of Defense, or the U.S. Government.

References

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

Gobet, F., Richman, H. B., Staszewski, J., and Simon, H. A. (1997).Goals, representations, and strategies in a concept formation task: The EPAM model. In *The Psychology of Learning and Motivation, Vol. 37*, D. L. Medin (Ed.), pp. 265-290. San Diego, CA: Academic Press.

Medin, D. L., & Smith, E. E. (1981). Strategies and classification learning. *Journal of Experimental Psychology: Human Learning and Memory, 7*(4), 241-253.

Newell, A. & Simon, H. A. (1972). *Human problem solving*. Englewood Cliffs, NJ: Prentice-Hall.

Nosofsky, R. M., Palmeri, T. J., &McKinley, S. C. (1994). Rule-plus-exception model of category learning. *Psychological Review*, *101*, 53-79.

Siegler, R. S. (1987). The perils of averaging data over strategies: An example from children's addition. *Journal of Experimental Psychology: General, 116*(3), 250-264.

Smith, J. D., & Minda, J. P. (2000). Thirty categorization results in search of a model. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 26*(1), 3-27.