Classification errors and response times over multiple distributed sessions as a function of category structure

Derek E. Zeigler^{1,2} · Ronaldo Vigo^{2,3}

Published online: 11 May 2018
© Psychonomic Society, Inc. 2018

Abstract

Learning difficulty orderings for categorical stimuli have long provided an empirical foundation for concept learning and categorization research. The conventional approach seeks to determine learning difficulty orderings in terms of mean classification accuracy. However, it is relatively rare that the stability of such orderings is tested over a period of extended learning. Further, research rarely explores dependent variables beyond classification accuracy that may also indicate relative learning difficulty, such as classification response times (RTs). Using a family of category structures defined over three binary dimensions and four positive examples that is well-known for its robust learning difficulty ordering, we report the results of two experiments that test the stability of the ordering (in terms of both errors and RTs) over multiple category learning sessions. The experimental stimuli consisted of instantiations of each of the six category structures in the family. These take the form of categories consisting of four “flasks” that vary along the binary features of size (large or small), shape (circular or triangular), and color (black or white). Experiment 1 shows that when participants are randomly presented instances of all six types, the difficulty ordering remains stable across all three sessions. This stability is present in terms of mean accuracy (errors) as well as mean RTs. In Experiment 2, participants were repeatedly exposed to category instances of a single type. In terms of errors, the ordering is revealed in the first session and disappears in later sessions. The opposite trend is observed for classification RTs: The ordering is not present in the first session but is revealed in later sessions. This suggests that even when individuals reach a relative degree of expertise in terms of reduced errors, the original degree of difficulty continues to influence processing. We interpret these results in the context of the concept learning and perceptual expertise literatures.

Keywords Concept learning · Categorization · Perceptual fluency · Expertise

The way in which category specific knowledge is acquired over time has been a long-standing central topic in the cognitive and perceptual sciences. Accordingly, the influence of training and experience on learning visual categories has been the focus of much empirical work (e.g., Gauthier & Tarr, 1997; Scott, Tanaka, Sheinberg, & Curran, 2008; Tanaka, Curran, & Sheinberg, 2005). This research often involves accounting for the results of concept learning tasks that necessitate classifying category members and nonmembers. Studies

in this area explore questions like the following: Can different concepts be ordered by their relative learning difficulty? Does repeated exposure to a concept result in perceptual expertise and/or expert classification? Is concept acquisition inherently easier for some individuals? Work in this area has explored whether different concepts can be ordered by their relative learning difficulty (e.g., Feldman, 2000; Kruschke, 1992; Nosofsky, Gluck, Palmeri, McKinley, & Gauthier, 1994; Shepard, Hovland, & Jenkins, 1961; Vigo, 2013). Indeed, for certain categories of objects (i.e., categorical stimuli), research has shown that robust learning difficulty orderings can be established in terms of how accurately subjects classify the objects as members of the category (Kruschke, 1992; Nosofsky, Gluck, et al., 1994; Vigo, 2013, 2014; Vigo, Evans, & Owens, 2015; Vigo, Zeigler, & Halsey, 2013). Such relative difficulty between categories tells us something fundamental about our conceptual system. Notably, relative classification accuracy can be used to reveal which relational

✉ Derek E. Zeigler
zeigler.d@wwc.edu

¹ Department of Psychology, West Virginia Wesleyan College, Buckhannon, WV 26201, USA

² Consortium for the Advancement of Cognitive Science, Ohio University, Athens, OH, USA

³ Department of Psychology, Ohio University, Athens, OH, USA

structures humans are most sensitive to in their environment. Sensitivity to certain structures likely results in better classification of its members, whereas categorical stimuli with structures for which we are relatively less sensitive likely results in a corresponding decrease in classification accuracy. In the current work, we explore the relationship between classification accuracy and classification response time when categories are learned over an extended period of time. Some research suggests that response times do not always reflect an individual's level of expertise in terms of accuracy (e.g., Chi, Feltovich, & Glaser, 1981; Rubenstein, 2013). Indeed, it is easy to see how different strategies for making a correct response may lead to the same performance in terms of accuracy but markedly different performance in terms of response time. In what follows, we explore the possibility that the response times will differ from error rates when making classification decisions about object stimuli. It is important for models of concept learning and perceptual expertise to account for this relationship between accuracy and response time because the type of strategy used during a task may provide researchers with important clues about the nature of underlying cognitive mechanisms during learning.

Concept learning difficulty orderings for categorical stimuli form much of the empirical foundation of concept learning research. As suggested, learning difficulty orderings are revealed by the relative classification error rates for different categories in a classification task. We now define some terms that will become critical in the following discussion. These definitions are similarly presented by Vigo (2014, p. 15) and Murphy (2002). First, we must distinguish between “concepts” and “categories.” In short, concepts are the internal mental representations of categories in the world. That is, concepts are assumed to correspond to (or represent) objective categories of objects. These categories are said to be “dimensionally definable” if the set of objects can be characterized in terms of a shared number of fixed dimensions. Moreover, we will refer to such a dimensionally definable set of object stimuli as “categorical stimuli.” We operationalize what will be referred to as “concept learnability” or “concept difficulty” in terms of categorization/classification performance. That is, the observable behaviors under consideration will be the proportion of classification errors and the response times (RTs) for classification decisions. The discussion of specific experiments to come will likely serve to clarify this terminology.

Models of conceptual behavior are generally tested in terms of classification performance. Specifically, researchers may obtain average learning times (time taken to learn a concept) or average learning accuracies (percentage of correct responses) in terms of correct object classifications for categories of varying learnability, and then fit model predictions to the aggregate response data across participants. Research with learning times typically uses a prespecified learning criterion

and learning proceeds until the criterion is met. In these tasks, the number of trials/blocks to criterion is the primary dependent variable (e.g., Nosofsky, Gluck, et al., 1994). Alternatively, research using learning accuracies typically takes mean proportion of errors as the primary dependent variable (e.g., Feldman, 2000; Vigo, 2013; Vigo et al., 2013). Additionally, the process by which concepts are learned can be studied by obtaining the rate of learning throughout an experimental session, and models can be subsequently compared by quantitatively predicting block-by-block changes in performance, providing a more comprehensive comparison than measures of aggregate performance. This approach has greatly advanced our understanding of concept formation.

Because of their robustness, concept learning difficulty orderings provide a stable baseline for the type of comparative studies between experts and novices that are involved in expertise acquisition (Chi et al., 1981; Tanaka & Taylor, 1991). They help tease apart more essential aspects of learning (e.g., structural aspects) from aspects that may depend more on qualitative preferences or long-term memory store. Indeed, the experimental approach taken here is motivated by previous work suggesting that relational cognition and pattern detection may play an important role in gaining expertise in particular categorization tasks (Kroger, Holyoak, & Hummel, 2004; Vigo, 2013, 2014).

Classification performance for sets of eight object stimuli that vary along three binary-valued separable dimensions has been the focus of much empirical work (Kruschke, 1992; Kurtz, Levering, Stanton, Romero, & Morris, 2013; Nosofsky, Gluck, et al., 1994; Rehder & Hoffman, 2005; Shepard et al., 1961; Vigo, 2013; Vigo et al., 2013). In classification tasks, this set is divided into two categories of four objects. For the eight possible unique stimuli, there are 70 distinct ways to partition them into two groups of four. These 70 partitions result in six structurally equivalent types in that any two instances of a type can be transformed into one another as a result of dimensional reassignment. To clarify, these instances are composed of four objects, and these instances are structurally equivalent in that they have the same category structure description. The 70 partitions are distributed among the six structurally equivalent types as such: I (6); II (6); III (24); IV (8); V (24); VI (2). Henceforth, we shall abbreviate this category structure as $3_2[4]$, using notation introduced by Vigo (2013, 2014). The first number (i.e., 3) stands for the number of dimensions; the index number (i.e., 2) stands for the arity of the dimensions (in this example, binary); and the third number (i.e., 4) stands for the number of object stimuli associated with the structure.

When participants are asked to learn the six structural relationships that can be partitioned from the $3_2[4]$ family, the following learning difficulty ordering emerges for the six structures: $I < II < [III, IV, V] < VI$, where Type I is the least

difficult to learn, followed by Type II. Types III, IV, and V are approximately equivalent in difficulty, and Type VI is the most difficult. An “instance” of a structure type refers to a categorical stimulus that is defined by the structure type.

The family of $3_2[4]$ structure types is of particular interest because of the dimensional size and cardinality of the categories. That is, three dimensions defining four positive examples turns out to be quite tractable for humans. This tractability seems to be due in part to the $3_2[4]$ family’s cardinality being within the limited boundaries of short-term memory capacity, which is now believed to be around four chunks of information (Cowan, 2001; Luck & Vogel, 1997). Accordingly, learning a category of four objects, each of which are composed of three dimensions, seems an appropriate test of category learning. Because of this privileged dimensional configuration, the error rates for these category types have been interpreted as revealing key psychological principles underlying the process of concept formation, including how object features influence category learning (Love, Medin, & Gureckis, 2004; Nosofsky, Palmeri, & McKinley, 1994; Vigo, 2009, 2011, 2013) as well as how observers distribute attention along object features (Kruschke, 1992; Nosofsky, 1984; Rehder & Hoffman, 2005). As such, the learning difficulty ordering for this category family serves as a particularly relevant benchmark for models of categorization behavior and concept formation.

A convenient method of showing category membership for categorical stimuli involves the use of Boolean cubes (e.g., Feldman, 2003). Members and nonmembers can be visualized in the form of a cube when stimulus dimensions take the value of 1 or 0 (i.e., true or false) in a D -dimensional space. Figure 1 depicts each of the $3_2[4]$ structures using Boolean cubes. Sides of the cube in Fig. 1 represent dimensions. Also, corners with circles represent positive examples and empty corners represent negative examples of the category. The thicker gray lines are meant to show simple relations between object stimuli. Notice the four simple relations in Type I, whereas Type VI has no simple relations.

Figure 2 shows a Boolean cube using example stimuli where the sides of the cube represent the dimensions of size, shape, and color. Corresponding dimensional changes occur when shifting along the sides of the cube. Figure 2 also shows example instances of each category type using the stimuli depicted in the cube. These flask-like stimuli are used in the current experiments and have been used in several previous studies (Vigo, 2013, 2014; Vigo, Doan, Doan, & Pinegar, 2018; Vigo et al., 2015; Vigo et al., 2013). Using stimuli that have been used in previous work provides the benefit of a common frame of reference in which meaningful comparisons can be made between studies. For instance, if our obtained classification data differs from previous work, we need not be concerned with the possibility that the chosen object stimuli resulted in differential processing. In short, we can expect

participants to process and classify these object stimuli in a way that is consistent with previous work. It should be noted that the $3_2[4]$ structures have also been studied using different types of stimuli, including insect stimuli (Vigo et al., 2015) and clock stimuli (Doan & Vigo, 2016; Vigo & Doan, 2015). In addition, they have been studied using two distinct empirical paradigms: a serioinformative and a parainformative paradigm (Vigo, 2013). Under the parainformative paradigm, all members of the category from which a concept is learned are presented simultaneously (i.e., in parallel). In contrast, under the serioinformative paradigm, the members of the category are presented one at a time (i.e., serially).

Shepard et al. (1961) were the first to establish the learning difficulty ordering for the $3_2[4]$ structures in a series of experiments in which classification tasks were preceded by identification tasks where unique responses were learned for each stimulus. For instance, Shepard et al. used a serioinformative task with corrective feedback after each classification in Experiment 1. Subjects classified members and nonmembers of categorical stimuli until a learning criterion of 32 consecutive correct responses was obtained. All participants were tested on Types I, II, and VI, but on only one of either Types III, IV, & V, resulting in a total of four types per subject. Five consecutive instances of each of the four types were presented to participants. Spatially separated stimuli were presented, and features varied in that they were either a screw or a bolt, a candle or light bulb, and either a violin or trumpet. The ordering was also observed using geometric objects in Experiment 2.

This learning difficulty ordering has remained relevant because of multiple subsequent replications. For instance, these structures were studied further by Nosofsky, Gluck, et al. (1994), who replicated the results obtained by Shepard et al. (1961) while providing block-by-block learning curves suitable for quantitative fitting. The resulting learning curves conform to the difficulty ordering in terms of percentage of correct responses, showing a high level representation of the difficulty of these types in terms of learning category membership across time. Accordingly, models of classification learning were able to be compared by fitting their predictions to these data (ALCOVE, Kruschke, 1992; the rational model, Anderson, 1991; the configural-cue model, Gluck & Bower, 1988).

In addition to these replications of serioinformative tasks, the ordering has also been replicated using parainformative tasks (Vigo, 2013; Vigo et al., 2013), where participants are able to study all category members (and nonmembers) in a learning phase before classifying objects one at a time.¹ Table 1 shows the reported mean classification errors per type

¹ Note that it is sufficient for all the members of the category to be displayed simultaneously for a concept learning task to be parainformative; that is, it is not necessary that the nonmembers also be displayed.

Fig. 1 Boolean cubes for the $3_2[4]$ structure types. Members of the $3_2[4]$ structure types are denoted by the corners of a cube. Sides of the cube represent dimensions. Corners with circles represent positive examples,

and empty corners represent negative examples of the category. The gray lines show simple relations between object stimuli

in studies that used a parainformative task without feedback. In all three tasks, participants studied category members and nonmembers in a 20-second learning phase. Following the

learning phase, participants classified serially presented stimuli. Vigo et al. (2015) gathered classification performance from adults, adolescents, and adolescents with ADHD. The

Fig. 2 Sample stimuli depicting the $3_2[4]$ structure types. The Boolean cube depicted is shown using stimuli from the current study. Sides of the cube represent the size, shape, and color. Moving along the sides of the cube results in a corresponding dimensional change. Example instances

of each category type are given for these dimensions. These are the same stimuli used in a number of previous studies (Vigo, 2013, 2014; Vigo et al., 2015; Vigo et al., 2013)

Table 1 Reported mean classification errors for $3_2[4]$ types in studies using a parainformative task without feedback

	Mean classification errors					
	I	II	III	IV	V	VI
Vigo (2013)	.02	.07	.19	.18	.17	.32
Vigo et al. (2013)	.01	.13	.19	.18	.19	.32
Vigo et al. (2015)	.02	.12	.17	.19	.20	.30

data presented in Table 1 correspond to the adult sample. These studies are of particular interest because the current work uses the same experimental design.

Additionally, Vigo (2013, Experiment 3) has demonstrated that upon a brief visual inspection, participants intuitively rank the difficulty of these structures in accordance with the empirical ordering in terms of correct classifications. Participants gave judgments of learning difficulty on a scale of 1 through 10 for instances of the $3_2[4]$ structures, which were simultaneously presented on each trial. Mean magnitude judgments for each of the six $3_2[4]$ structures were reported as follows: I (1.7) < II (4.3) < [III (5.5), IV (5.6), V (5.5)] < VI (6.3). Vigo (2013) interprets this result as evidence for the capacity to form metacognitive judgments about concept learnability. Further, such metacognitive judgments are accurate in terms of classification performance.

It should be noted that there are studies where the classic ordering was not observed (Feldman, 2000; Love & Markman, 2003; Lewandowsky, 2011; Love, 2002; Nosofsky & Palmeri, 1996; Rehder & Hoffman, 2005) leading Kurtz et al. (2013) to argue that the classic Type II advantage over Type IV is achieved as a result of (a) the particular interplay between stimulus material and dimensional assignment, and (b) instructional language that places emphasis on “rules” and “features.” However, the alternative orderings are likely due to experimental designs that do not sample from all possible structure instances. The learning difficulty of these types is meant to be an aggregated average proportion of performance across all instances of the type, and a limited subset of these instances will quite likely result in deviations from the true difficulty ordering.

In summary, classification performance on the six $3_2[4]$ structure types is significant because it provides an empirically validated difficulty ordering by which concept learning models can be compared. These results gain new significance in light of recent developments in the way that few models can predict the ordering precisely (GCM, Nosofsky, 1984; ALCOVE, Kruschke, 1992; SUSTAIN, Love et al., 2004; GISTM-SE, Vigo, 2013) and recent accounts (Feldman, 2000; Goodwin & Johnson-Laird, 2011; Kemp, 2012; Vigo, 2006) do not predict the learning difficulty ordering.

However, it is rare that the stability of such orderings is tested on classification performances as it unfolds over time. Further, this research rarely explores dependent variables beyond classification accuracy, which may also indicate relative learning difficulty.

Accordingly, the current work reports the results of a category learning task testing classification errors and response times (RTs) on the $3_2[4]$ structures over multiple experimental sessions taking place on consecutive days. Many researchers (see Palmeri & Cottrell, 2009) operationalize the mechanisms of expertise as a normative outcome of the learning process. Accordingly, observing classification over time in the current experiments allows us to consider categorization processes as a domain of perceptual expertise. The development of expertise goes considerably beyond the accumulation of vast amounts of knowledge. In addition to the learning effects that commonly result from experience or training, more complex and accurate perceptual representations also emerge over time. That is, increased task performance and more accurate perceptual representations are typically associated with a concurrent effect of processing fluency marked by quicker response times as well as a reduction in the cognitive and attentional resources dedicated to the task. However, it should be noted that in expertise experiments, reaction time data often do not reflect the individual’s degree of expertise on a problem-solving task (Chi et al., 1981). Indeed, from the standpoint of error rates, people can often perform with similar accuracy on two distinct problem-solving tasks while resorting to different strategies (implicitly or explicitly) in solving the problems. The type of strategy used can provide important clues as to the nature of its underlying cognitive mechanisms. This leaves open the possibility that the RTs will differ from the error rates. Additionally, there is a debate in the decision-making litera-

Fig. 3 Learning phase for an instance of Type I in the current experiment

ture whether quicker RTs lead to errors in decision-making. It seems that under certain circumstances, quicker RTs lead to more errors, whereas in other circumstances quicker RTs lead to fewer errors. Indeed, there is no correlation between RTs and accuracy of decisions when sampling across a number of different decision-making problems (Rubenstein, 2013). Furthermore, as rigorous cognitive scientists, we are interested in establishing the robustness of concept learning difficulty orderings rather than relying on intuition. We believe that the determination of learning difficulty orderings from the standpoint of RTs is an important and neglected step in achieving this goal.

We also propose that if concept difficulty is ultimately determined by concept structure and that no amount of training can change the tested difficulty orderings, expertise must be domain specific and not structure specific. This would mean that what best characterizes experts is the nature of the dimensional spaces they learn and not primarily the nature of the concept structures they learn.

Accordingly, the current work repeatedly exposes participants to category instances in order to obtain measures of both classification accuracy as well as classification response times over time. This is significant because it is possible that the $3_2[4]$ learning difficulty might only be obtained when categorical stimuli are novel. It has remained an empirical question whether the $3_2[4]$ ordering is observed after the categorical stimuli are no longer novel.

The current experiments

Experiments 1 and 2 both test $3_2[4]$ category structures so that the learning difficulty ordering can be assessed across sessions. Further, both studies involve category learning tasks in which classification errors and response times (RTs) are obtained over three experimental sessions taking place on 3 consecutive days. This procedure is motivated by the design of previous research exploring the influence of experience/training on category learning. Precedence is shown for testing participants on consecutive days (Nosofsky, Little, Donkin, & Fific, 2011; Nosofsky & Palmeri, 1997; Nosofsky & Stanton, 2005). For instance, Nosofsky and Palmeri (1997) and Nosofsky et al. (2011) both tested participants over 5 consecutive days. Such studies offer more benefits than experiments that use multiple sessions on a single day (Palmeri, 1997) in that they better resemble real-world experiential learning, which typically requires repeated exposure to domain-specific stimuli (Myles-Worsley, Johnston, & Simons, 1988).

In Experiment 1, each participant is exposed to all six structure types of the $3_2[4]$ category family in order to determine if the learning difficulty ordering remains stable over multiple

experimental sessions. The stability of the ordering is tested both in terms of proportion of classification errors and classification response times (RTs). Testing multiple category structures per person allows us to estimate parameters at an individual level.

In Experiment 2, each participant is exposed to a single structure type of the $3_2[4]$ category family. The obtained results are compared with those of Experiment 1 to determine if the relationship between errors, RTs, and experience varies according to the number of category structures the participant must learn.

Experiment 1

Method

Stimuli and materials Our stimuli consist of flasks that vary on the binary features of size (large or small), shape (circular or triangular), and color (black or white), in accordance with the structure types associated with the $3_2[4]$ category family. These stimuli are the same stimuli used by Vigo (2013). Participants were tested on instances associated with all six category structures of the $3_2[4]$ family. An HP XW4600 workstation with a Dell 1708FP 15-in. flat-panel LCD monitor (5-ms response time) was used to display the stimuli.

Participants Forty participants were tested on the six category structures. Participants completed three sessions on 3 consecutive days. All participants were undergraduate students enrolled in an introductory psychology course at Ohio University.

Procedure Participants engaged in a classification task over the course of three experimental sessions on 3 consecutive days. In Experiment 1, participants gave eight responses per block across a 36-block session, resulting in 288 responses per session. Three 36-block sessions resulted in a total of 108 total blocks per subject over the 3 days. The classification task is unsupervised (without feedback) and parainformative in nature (Vigo, 2013, 2014), meaning that category members are presented in a learning phase prior to classification. The learning phase consists of two sets of bottles spatially separated by a line in the middle of the screen. The four category members are presented above the line while the four nonmembers are presented below the line, and the locations of members and nonmembers are randomized across trials. Figure 3 shows an example of what participants saw during the learning phase. Before the experiment began, participants were told that each display represents the preferences of a bottle collector and that their task was to learn which bottles the collector likes.

Participants were told that the four “liked” bottles were presented above a line on the screen and the four “not liked” bottles were presented below the line on the screen. After 20 s of studying the category members and nonmembers in the learning phase, individual flasks were serially presented for classification. Subjects were told that their task was to classify each individual bottle as either “liked” or “not liked” using the mouse: a left mouse click (labelled “Y”) meant the bottle was “liked,” and a right mouse click (labelled “N”) meant the bottle was “not liked” by the bottle collector. After classification responses were given for the eight bottles, a fixation cross appeared in the center of the screen, and the next category block was shown. The fixation cross was presented for 1 second before the next learning phase was presented. Participants completed three sessions on 3 consecutive days. Six instances of each structure type were presented randomly within each session, resulting in a 36-block session lasting approximately 30 minutes each. Structure instances, as well as the eight objects presented for categorization, were presented to participants randomly throughout the experimental session in order to control for order effects.

Classification accuracy and response times were recorded on each experimental trial. Response times were excluded from the analysis if $RT < 0.025$ seconds. Average response times per block were excluded if more than half of the trial RTs were missing. That is, at least four of the eight RTs per block were needed to be included in the analysis.

Results

Classification errors Mean proportion errors per session for each of the six structure types are shown in Fig. 4a. Mean proportion error, standard deviation, standard errors, and 95% confidence interval for each type in Session 1 are as follows: Type I ($M = 0.01$, $SD = 0.01$, $SE = 0.002$, $CI [0.009, 0.011]$); Type II ($M = 0.10$, $SD = 0.099$, $SE = 0.016$, $CI [0.095, 0.105]$); Type III ($M = 0.17$, $SD = 0.076$, $SE = 0.012$, $CI [0.166, 0.174]$); Type IV ($M = 0.18$, $SD = 0.091$, $SE = 0.014$, $CI [0.176, 0.184]$); Type V ($M = 0.18$, $SD = 0.101$, $SE = 0.016$, $CI [0.175, 0.185]$); Type VI ($M = 0.28$, $SD = 0.132$, $SE = 0.021$, $CI [0.273, 0.287]$). For Session 2, mean proportion error, standard deviation, and 95% confidence interval for each type are as follows: Type I ($M = 0.01$, $SD = 0.027$, $SE = 0.004$, $CI [0.009, 0.011]$); Type II ($M = 0.05$, $SD = 0.069$, $SE = 0.011$, $CI [0.047, 0.053]$); Type III ($M = 0.10$, $SD = 0.086$, $SE = 0.014$, $CI [0.096, 0.104]$); Type IV ($M = 0.13$, $SD = 0.096$, $SE = 0.015$, $CI [0.125, 0.135]$); Type V ($M = 0.13$, $SD = 0.096$, $SE = 0.015$, $CI [0.125, 0.135]$); Type VI ($M = 0.21$, $SD = 0.136$, $SE = 0.022$, $CI [0.203, 0.217]$). For Session 3, mean proportion error, standard deviation, and 95% confidence interval for each type are as follows: Type I ($M = 0.01$, $SD = 0.019$, $SE = 0.003$, $CI [0.009, 0.011]$); Type II ($M = 0.05$, $SD = 0.08$, $SE = 0.01$, $CI [0.047, 0.053]$); Type III ($M = 0.09$, $SD = 0.077$, $SE = 0.012$, $CI [0.086, 0.094]$); Type IV ($M = 0.11$, $SD = 0.08$, $SE = 0.013$, $CI [0.106, 0.114]$); Type V ($M = 0.11$, $SD = 0.094$, $SE = 0.015$, $CI [0.105, 0.115]$); Type VI ($M = 0.17$, $SD = 0.158$, $SE = 0.025$, $CI [0.162, 0.178]$).

Fig. 4 **a** Mean proportion of classification errors for each of the six structure types within the $3_2[4]$ category family across all three sessions of Experiment 1. Mean classification error values and standard error bars are presented for each column. **b** Mean response times (RTs) for each of

the six structure types within the $3_2[4]$ category family across all three sessions of Experiment 1. Mean RT values (in milliseconds) and standard error bars are presented for each column. Roman numerals indicate the learning difficulty ordering according to pairwise significance tests

A two-way repeated-measures analysis of variance (ANOVA) on the obtained proportion of errors, using session and structure type as within-subjects variables, revealed a significant main effect for both session, $F(2, 78) = 37.64$, $p < .001$, $\eta_p^2 = .491$, and structure type, $F(5, 195) = 58.40$, $p < .001$, $\eta_p^2 = .600$. Additionally, there was a significant interaction between session and type, $F(10, 390) = 4.50$, $p < .001$, $\eta_p^2 = .103$. In all sessions, pairwise t tests show that the mean proportion errors per type conform to the classic difficulty ordering, $I < II < [III, IV, V] < VI$.

Paired samples t tests reveal significant reductions in mean errors between Session 1 and Session 2 for all structure types, except Type I, which showed marginal significance, $t(39) = -1.706$, $p = .096$, Cohen's $d = 2.33$. Conversely, paired-samples t tests revealed that mean errors between Session 2 and Session 3 were nonsignificant for all types.

Classification response times Mean response times (RTs) per session for each of the six structure types are shown in Fig. 4b. Mean RTs (ms), standard deviation, and 95% confidence interval for each type in Session 1 are as follows: Type I ($M = 348$, $SD = 108.26$, $SE = 17.12$, CI [342.69, 353.31]); Type II ($M = 827$, $SD = 221.57$, $SE = 35.03$, CI [816.14, 837.86]); Type III ($M = 932$, $SD = 229.32$, $SE = 36.26$, CI [920.76, 943.24]); Type IV ($M = 904$, $SD = 275.92$, $SE = 43.63$, CI [890.48, 917.52]); Type V ($M = 900$, $SD = 274.10$, $SE = 43.34$, CI [886.57, 913.43]); Type VI ($M = 1066$, $SD = 273.63$, $SE = 43.23$, CI [1052.60, 1079.40]). For Session 2, Mean RTs (ms), standard deviation, and 95% confidence interval for each type are as follows: Type I ($M = 311$, $SD = 87.29$, $SE = 13.8$, CI [306.72, 315.28]); Type II ($M = 710$, $SD = 150.40$, $SE = 23.78$, CI [702.63, 717.37]); Type III ($M = 807$, $SD = 207.99$, $SE = 32.89$, CI [796.81, 817.19]); Type IV ($M = 840$, $SD = 208.80$, $SE = 33.01$, CI [829.77, 850.23]); Type V ($M = 783$, $SD = 176.64$, $SE = 27.93$, CI [774.34, 791.66]); Type VI ($M = 986$, $SD = 251.46$, $SE = 39.77$, CI [973.68, 998.32]). For Session 3, Mean RTs (ms), standard deviation, and 95% confidence interval for each type are as follows: Type I ($M = 278$, $SD = 64.63$, $SE = 10.22$, CI [274.83, 281.17]); Type II ($M = 604$, $SD = 132.32$, $SE = 20.92$, CI [597.52, 610.48]); Type III ($M = 725$, $SD = 187.88$, $SE = 29.71$, CI [715.79, 734.21]); Type IV ($M = 763$, $SD = 182.97$, $SE = 28.93$, CI [754.03, 771.97]); Type V ($M = 722$, $SD = 169.65$, $SE = 26.82$, CI [713.69, 730.31]); Type VI ($M = 931$, $SD = 242.77$, $SE = 38.39$, CI [919.1, 942.9]).

A two-way repeated-measures ANOVA on the obtained RTs, using session and structure type as within-subjects variables, revealed a significant main effect for both session, $F(2, 78) = 33.94$, $p < .001$, $\eta_p^2 = 0.465$, and structure type, $F(5, 195) = 203.59$, $p < .001$, $\eta_p^2 = 0.839$. Additionally, there was a significant interaction between session and type, $F(10, 390)$

$= 3.13$, $p < .001$, $\eta_p^2 = 0.074$. In Session 1, paired-samples t tests show that the mean RTs per type conform to the classic difficulty ordering, $I < II < [III, IV, V] < VI$. Paired-samples t tests showed significant differences between Type IV and Type V structures in both Session 2, $t(39) = 2.646$, $p = .012$, Cohen's $d = 1.86$, and Session 3, $t(39) = 2.082$, $p = .044$, Cohen's $d = 1.47$. Other than these differences, all other paired-samples t tests conformed to the classic difficulty ordering in Sessions 2 and 3.

Between Session 1 and Session 2, paired-samples t tests reveal significant reductions in mean RTs across types, with the exception of marginal significance shown by Type IV, $t(39) = 1.81$, $p = .079$, Cohen's $d = 1.65$, and Type VI, $t(39) = 1.85$, $p = .073$, Cohen's $d = 1.93$. Between Session 2 and Session 3, paired-samples t tests reveal significant reductions in mean RTs for all types except Type VI, $t(39) = 1.73$, $p = .092$, Cohen's $d = 1.41$.

Recall that response times were excluded from the analysis if $RT < 0.025$ seconds. There was an average of 11.45 ($SD = 0.94$) excluded RTs per session in Experiment 1 (out of 288 responses per session; i.e., 36 blocks with eight responses each). Excluded response times did not differ significantly across sessions: Session 1 ($M = 12.35$, $SD = 1.49$); Session 2 ($M = 9.55$, $SD = 1.46$); Session 3 ($M = 12.45$, $SD = 1.89$); $F(2, 117) = 1.03$, $p = 0.36$. Average response times per block were excluded if more than half of the trial RTs were missing (five of the eight RTs per block). This only occurred in nine blocks across all participants in Experiment 1. The reported pattern of results remains the same if these data are not excluded from the analyses.

Order effects We looked for order effects in the data from Experiment 1. Specifically, we looked for cases where Type II and Type VI are presented back-to-back to participants. We chose Type II and Type VI for this analysis because we believe that if any order effects exist they would exist between these two types. This is because they are the extreme cases, where Type II is nontrivial yet relatively easy, and Type VI is difficult. In other words, a simple rule can be formed for Type II, and there is no simple rule for Type VI. Across all 40 participants, there were 147 instances of Type II directly proceeding Type VI and 142 instances where Type VI directly preceded Type II. In terms of average errors, Type VI seems to become a bit easier when it is presented after Type II ($M = 0.19$, $SD = 0.02$) as compared with when it is presented before Type II ($M = 0.23$, $SD = 0.02$). However, these means are nonsignificant, $t(281) = -1.40$, $p = .08$. There was no significant difference in errors for Type II whether it was presented before ($M = 0.07$, $SD = 0.01$) or after ($M = 0.07$, $SD = 0.01$) Type VI, $t(283) = 1.65$, $p = .969$, Cohen's $d = 0.07$. Additionally, we looked for order effects in the response time data for these two types. There was no significant difference in response times (ms)

for Type II whether it was presented before ($M = 720$, $SD = 20.86$) or after ($M = 697$, $SD = 21.87$) Type VI, $t(286) = 0.77$, $p = .22$, Cohen's $d = 1.08$. Likewise, there was no significant difference in response times (ms) for Type VI whether it was presented before ($M = 1004$, $SD = 32.79$) or after ($M = 1017$, $SD = 29.39$) Type II, $t(282) = 0.30$, $p = .38$, Cohen's $d = 0.42$.

Discussion

The reported data from Experiment 1 suggests that if participants are randomly presented instances of all six structure types, the difficulty ordering remains stable across all three sessions. This stability is present in terms of mean accuracy (errors) and nearly shown in terms of mean response times (only marginal p values between Types IV and V create deviations from the ordering).

Recall that learning difficulty orderings for categorical stimuli have long provided an empirical foundation for concept learning and categorization research (e.g., Feldman, 2000; Nosofsky, Gluck, et al., 1994; Shepard et al., 1961; Vigo, 2013; Vigo et al., 2015; Vigo et al., 2013). The current experiments were designed to extend this empirical work in two critical areas. First, the stability of the $3_2[4]$ difficulty ordering was tested by obtaining classification data over multiple sessions. Second, classification RTs were treated as a measure of concept learning difficulty by considering the relative RT orderings. The conventional approach in this area of research has sought to establish difficulty orderings in terms of mean classification accuracy (proportion of correct/incorrect classification responses) from a single learning session. We are unaware of a study testing the stability of such orderings over a period of extended learning. The observed results of Experiment 1 seem to be the first to show that this learning difficulty ordering remains stable over time and across multiple learning sessions. This is significant because it is possible that the $3_2[4]$ learning difficulty might only be obtained when categorical stimuli are novel. It has remained an empirical question whether the ordering remains after the categorical stimuli are no longer novel. The stability across sessions of Experiment 1 provides further support for the robust learning difficulty ordering associated with the $3_2[4]$ structures. Indeed, given the number of replications of the ordering, one might consider it an empirical law of human classification learning.

Experiment 2

Method

Stimuli and materials As in Experiment 1, the stimuli are flasks that vary on the binary features of size (large or small), shape (circular or triangular), and color (black or white) in

accordance with the structure types associated with the $3_2[4]$ category family. Participants were tested on instances associated with one of the six category structures of the $3_2[4]$ family. An HP XW4600 workstation with a Dell 1708FP 15-in. flat-panel LCD monitor (5-ms response time) was used to display the stimuli.

Participants A total of 180 participants were tested on one of the six category structures ($n = 30$ per structure), described above. Participants completed three sessions on 3 consecutive days. All participants were undergraduate students enrolled in an introductory psychology course at Ohio University.

Procedure Participants engaged in a classification task over the course of three experimental sessions on 3 consecutive days. The procedure in Experiment 2 is nearly identical to the procedure used in Experiment 1. In Experiment 2, participants gave eight responses per block across a 24-block session, resulting in 192 responses per session. Three 24-block sessions resulted in a total of 72 total blocks per subject over the 3 days.

Results

Classification errors Mean proportion errors per session for each of the six structure types are shown in Fig. 5a. Mean proportion error, standard deviation, and 95% confidence interval for each type in Session 1 are as follows: Type I ($M = 0.03$, $SD = 0.019$, $SE = 0.003$, CI [0.029, 0.031]); Type II ($M = 0.09$, $SD = 0.071$, $SE = 0.01$, CI [0.086, 0.094]); Type III ($M = 0.18$, $SD = 0.10$, $SE = 0.02$, CI [0.173, 0.187]); Type IV ($M = 0.20$, $SD = 0.085$, $SE = 0.02$, CI [0.193, 0.207]); Type V ($M = 0.21$, $SD = 0.102$, $SE = 0.02$, CI [0.203, 0.217]); Type VI ($M = 0.28$, $SD = 0.124$, $SE = 0.02$, CI [0.273, 0.287]). For Session 2, the mean proportion error, standard deviation, and 95% confidence interval for each type are as follows: Type I ($M = 0.01$, $SD = 0.007$, $SE = 0.001$, CI [0.0096, 0.0104]); Type II ($M = 0.03$, $SD = 0.029$, $SE = 0.005$, CI [0.028, 0.032]); Type III ($M = 0.12$, $SD = 0.105$, $SE = 0.019$, CI [0.113, 0.127]); Type IV ($M = 0.13$, $SD = 0.08$, $SE = 0.015$, CI [0.125, 0.135]); Type V ($M = 0.14$, $SD = 0.099$, $SE = 0.018$, CI [0.134, 0.146]); Type VI ($M = 0.14$, $SD = 0.123$, $SE = 0.023$, CI [0.132, 0.148]). For Session 3, the mean proportion error, standard deviation, and 95% confidence interval for each type are as follows: Type I ($M = 0.004$, $SD = 0.005$, $SE = 0.001$, CI [0.0036, 0.0044]); Type II ($M = 0.03$, $SD = 0.028$, $SE = 0.005$, CI [0.028, 0.032]); Type III ($M = 0.08$, $SD = 0.0770.014$, CI [0.075, 0.085]); Type IV ($M = 0.10$, $SD = 0.084$, $SE = 0.015$, CI [0.095, 0.105]); Type V ($M = 0.10$, $SD = 0.079$, $SE = 0.015$, CI [0.095, 0.105]); Type VI ($M = 0.10$, $SD = 0.129$, $SE = 0.024$, CI [0.091, 0.109]).

Fig. 5 **a** Mean proportion of classification errors for each of the six structure types within the $3_2[4]$ category family across all three sessions of Experiment 2. Mean classification error values and standard error bars are presented for each column. Roman numerals indicate the learning difficulty ordering according to pairwise significance tests

the six structure types within the $3_2[4]$ category family across all three sessions of Experiment 2. Mean RT values (in milliseconds) and standard error bars are presented for each column. Roman numerals indicate the learning difficulty ordering according to pairwise significance tests

A repeated-measures ANOVA on the obtained proportion of errors, using session as a within-subjects variable and structure type as a between subjects variable, revealed a significant main effect for both session, $F(2, 348) = 183.83, p < .001, \eta_p^2 = .514$, and structure type, $F(5, 174) = 21.48, p < .001, \eta_p^2 = .382$. Additionally, there was a significant interaction between session and type, $F(10, 348) = 10.22, p < .001, \eta_p^2 = .227$. In Session 1, pairwise t tests show that the mean proportion of errors per type conform to the classic difficulty ordering, $I < II < [III, IV, V] < VI$. Pairwise t tests on the data from Sessions 2 and 3 reveal alternative difficulty orderings in terms of classification errors, $I < II < [III, IV, VI, V]$ and $I < II < [III, IV, V, VI]$, respectively.

Paired-samples t tests reveal significant reductions in mean errors between Session 1 and Session 2 for all structure types. Between Session 2 and Session 3, paired-samples t tests reveal significant reductions in mean errors for all types except Type I, $t(29) = 1.70, p = .536$, Cohen's $d = 1.16$, and Type II, $t(29) = 1.70, p = .249$, Cohen's $d = 1.36$.

Classification response time Mean response times (RTs) per session for each of the six structure types are shown in Fig. 5b. Mean RTs (ms), standard deviation, and 95% confidence interval for each type in Session 1 are as follows: Type I ($M = 366, SD = 27.7, SE = 5.06, CI [334.19, 337.81]$); Type II ($M = 804, SD = 149.05, SE = 27.21, CI [794.26, 813.74]$); Type III

($M = 819, SD = 176.96, SE = 32.31, CI [807.44, 830.56]$); Type IV ($M = 854, SD = 217.56, SE = 39.72, CI [839.79, 868.21]$); Type V ($M = 875, SD = 205.56, SE = 37.53, CI [861.57, 888.43]$); Type VI ($M = 1090, SD = 226.69, SE = 41.39, CI [1075.19, 1104.81]$). For Session 2, Mean RTs (ms), standard deviation, and 95% confidence interval for each type are as follows: Type I ($M = 229, SD = 21.2, SE = 3.87, CI [216.62, 219.38]$); Type II ($M = 645, SD = 166.66, SE = 30.43, CI [634.11, 655.89]$); Type III ($M = 709, SD = 174.56, SE = 31.87, CI [697.60, 720.40]$); Type IV ($M = 791, SD = 248.84, SE = 45.43, CI [774.74, 807.26]$); Type V ($M = 823, SD = 218.52, SE = 39.40, CI [808.9, 837.1]$); Type VI ($M = 970, SD = 226.98, SE = 41.44, CI [955.17, 984.83]$). For Session 3, Mean RTs (ms), standard deviation, and 95% confidence interval for each type are as follows: Type I ($M = 219, SD = 21.96, SE = 4.01, CI [215.57, 218.43]$); Type II ($M = 585, SD = 159.96, SE = 29.21, CI [574.55, 595.45]$); Type III ($M = 699, SD = 186.36, SE = 34.02, CI [686.83, 711.17]$); Type IV ($M = 726, SD = 273.74, SE = 49.98, CI [708.11, 743.89]$); Type V ($M = 763, SD = 228.01, SE = 41.63, CI [748.1, 777.9]$); Type VI ($M = 877, SD = 192.42, SE = 35.13, CI [864.43, 889.57]$).

A repeated-measures ANOVA on the obtained RTs, using session as a within-subjects variable and structure type as a between subjects variable, revealed a significant main effect for both session, $F(2, 348) = 104.66, p < .001, \eta_p^2 = .376$, and

structure type, $F(5, 174) = 62.4$, $p < .001$, $\eta_p^2 = .642$. Additionally, there was a significant interaction between session and type, $F(10, 348) = 2.56$, $p = .009$, $\eta_p^2 = .069$. In Session 1 and 2, pairwise t tests reveal alternative difficulty orderings in terms of classification RTs, $I < [II, III, IV, V] < VI$ and $I < [II, III]$, $[III, IV]$, $[IV, V] < VI$, respectively. In Session 3, however, pairwise t tests show that mean RTs per type conform to the classic difficulty ordering, $I < II < [III, IV, V] < VI$. Paired-samples t tests reveal significant reductions in mean RTs between Session 1 and Session 2 for all structure types. Between Session 2 and Session 3, paired-samples t tests reveal significant reductions in mean RTs for all types except Type I, $t(29) = 1.70$, $p = .543$, Cohen's $d = 0.37$, and Type III, $t(29) = 1.70$, $p = .646$, Cohen's $d = 0.32$.

Recall that response times were excluded from the analysis if $RT < 0.025$ seconds. There was an average of 4.98 ($SD = 0.33$) excluded RTs per session in Experiment 2 (out of 192 responses per session; i.e. 24 blocks with eight responses each). Unlike Experiment 1, excluded RTs differed significantly across sessions: Session 1 ($M = 6.41$, $SD = 0.56$); Session 2 ($M = 4.76$, $SD = 0.69$); Session 3 ($M = 3.75$, $SD = 0.41$); $F(2, 459) = 5.66$, $p < .01$. Post hoc comparisons indicate that the mean number of excluded RTs for Session 1 was significantly different than Session 3, $t(154) = 3.85$, $p < .01$, Cohen's $d = 5.40$. However, Session 2 did not significantly differ from Session 1, $t(154) = 1.86$, $p = .06$, Cohen's $d = 2.61$, or Session 3, $t(154) = 1.26$, $p = .21$, Cohen's $d = 1.78$. Mean number of excluded RTs also differed significantly across types: Type I ($M = 0.94$, $SD = 0.13$); Type II ($M = 3.04$, $SD = 0.44$); Type III ($M = 6.7$, $SD = 1.25$); Type IV ($M = 5.8$, $SD = 0.60$); Type V ($M = 4.94$, $SD = 0.50$); Type VI ($M = 4.91$, $SD = 0.63$); $F(5, 534) = 9.31$, $p < .001$. Post hoc tests reveal that mean number of excluded RTs for Type I was significantly different from all other types. Excluded RTs for Type II were also significant from all other types. All other pairwise tests were nonsignificant. Recall that average response times per block were excluded if more than half of the trial RTs were missing (five of the eight RTs per block). This only occurred in nine blocks across all participants in Experiment 2. The reported pattern of results remains the same if these data are not excluded from the analyses.

First six types analysis We sought to test whether decreases in errors were more likely based on (1) how many structure types participants were responsible for learning, or (2) the total number of structure instances encountered. Indeed, perhaps it was the total number of instances participants are being exposed to that is causing reduction in errors rather than the fact they are only being shown one structure type. Since each participant saw only six instances of each structure type in each session of Experiment 1, we thought that we could average across only

the first six instances of Experiment 2 instead of averaging across all 24 instances of each type. This would provide a comparison where number of instances per type are held constant across the experiments. If there is no decrease in errors in this new analysis as compared with Experiment 1, then being exposed to many instances must be what caused the observed decreases in Experiment 2. If there are decreases in this new analysis, then the decreases can confidently be attributed to only being exposed to one type. When only averaging across the first six types per session in Experiment 2, the mean proportion errors are actually much higher in Experiment 2, Session 1: I (.06), II (.18), III (.21), IV (.27), V (.25), VI (.39) relative to the reported mean proportion errors in Experiment 1 Session 1. However, in Sessions 2 and 3, we have a very similar pattern of results to when we are averaging across all 24 instances. Session 2: I (.01), II (.04), III (.11), IV (.14), V (.14), VI (.15); Session 3: I (.01), II (.02), III (.07), IV (.08), V (.09), VI (.11). The decrease in errors from Session 1 to Sessions 2 and 3 suggests that the decreases in errors in Experiment 2 might have a lot to do with just the sheer number of instances encountered. However, the drastic decreases in Sessions 2 and 3 (for Type VI in particular) suggest that only being exposed to one type is necessary for expertise in at least some circumstances.

Discussion

In Experiment 2, when participants are repeatedly exposed to category instances of a single type, the $3_2[4]$ ordering is observed in terms of errors in the first session, but not in later sessions. In terms of classification response times (RTs), the ordering is not present in the first session, but is revealed in later sessions.

These results suggest that the stability of the ordering shown in Experiment 1 (based on error rates) depends on the number of category structure types observers are exposed to during learning. When category structures are randomly sampled from all six types (as in Experiment 1), there is stability across sessions. However, when participants are given instances of a single type (as in Experiment 2) the ordering varies across sessions. The source of this instability likely stems from the relative degree of expertise participants are able to achieve when presented with so many instances of the same category type. The task of Experiment 2 learners seems relatively less demanding than that of Experiment 1 learners who are repeatedly exposed to category structures of varying degrees of complexity. Indeed, Experiment 2 learners provided more accurate classifications than Experiment 1 learners in terms of aggregate mean classification errors. Compare the respective error and RT data from both studies by examining Figs. 4 and 5. The mean proportions of errors are lower in Experiment 2 across every type

(with the exception of Type I). This result highlights the importance of how observers generalize under different sampling situations. Category learning experiments using categorical stimuli vary in terms of the sampling procedures used. For instance, Shepard et al. (1961) presented participants with all six structures, whereas Nosofsky, Gluck, et al. (1994) tested participants on only two of the six structures. Vigo and his collaborators (Vigo, 2014; Vigo et al., 2015; Vigo & Zeigler, 2016; Vigo et al., 2013) have employed consistent procedures when using the $3_2[4]$ structures, sampling from all six structure types. Previous studies that did not sample from all six structures still observed the ordering. The ordering was obtained because mean errors were taken based on the averaging across a single experimental session. The current findings suggest that if participants continued learning a limited set of category structures independently over the course of multiple sessions, we could expect the error ordering to disappear over time.

An additional empirical contribution of the current study concerns the obtained RT orderings and the treatment of classification RTs as a measure of concept learning difficulty. Previous research has not typically explored dependent variables other than classification accuracy. The results of Experiment 2 show that RTs might be a better measure of relative difficulty when measuring performance in an area where individuals have gained a lot of experience. These results suggest that as subjects' performance increases to perfect or near-perfect levels of classification, reaction times still reveal key differences in learning difficulty. This is a result that might be useful in many real-world applications. For instance, consider a field where experts rarely make mistakes. It may be difficult to study their performance in this area using measures of accuracy given the limited number of errors. Our results suggest that RTs can be used as a measure of task difficulty, allowing researchers the ability to order the relative difficulty of these tasks.

General discussion

In summary, Experiment 1 shows that when participants are randomly presented instances of all six $3_2[4]$ structure types, the difficulty ordering remains stable across all three sessions. This stability is present in terms of mean accuracy (errors) as well as mean RTs. In Experiment 2, participants were repeatedly exposed to category instances of a single type. In terms of errors, the ordering is revealed in the first session and disappears in later sessions. The opposite trend is observed for classification RTs: The ordering is not present in the first session but is revealed in later sessions.

When considering possible explanations for the different orderings between studies, it might be helpful to consider

previous research in the area of expertise. Work in this area has established the domain-specific nature of expertise (Bédard & Chi, 1992; Bukach, Phillips, & Gauthier, 2010; Searston & Tangen, 2017; Sims & Mayer, 2002; Voss, Greene, Post, & Penner, 1983). That is, expertise develops in the domain of application and does not typically generalize to other areas. For example, if experts with doctoral degrees are given problems outside of their subject, their performance is the same as undergraduate students (Voss et al., 1983). Additionally, expertise in the domain of fingerprint discrimination does not appear to generalize to other visual materials (Searston & Tangen, 2017), and expertise in the domain of classifying modern car models does not appear to translate to older car models (Bukach et al., 2010). Further, skilled Tetris players can show spatial expertise within the domain of Tetris-like shapes, but this does not appear to influence their performance on other tests of spatial ability (Sims & Mayer, 2002). In this context, the results of the current study seem to support and extend this line of work. Specifically, the relative gains in classification accuracy for Experiment 2 learners seems to support the domain-specific nature of categorizing these object stimuli. The better accuracy of Experiment 2 learners may have resulted simply due to receiving more overall instances of a given type than Experiment 1 learners. The implication is that conceptual expertise may develop acutely for specific categories in which we receive an abundance of experience.

Future work in this area is also necessary since this result seems to be at odds with a large body of research demonstrating that mixed or interleaved practice (vs. massed practice of one concept or category) often produces superior learning. Indeed, superior learning using mixed and interleaved practice has been shown specifically in the domain of visual categories (e.g., Hatala, Brooks, & Norman, 2003; Kang & Pashler, 2012; Kornell & Bjork, 2008). Based on these previous studies, we might expect overall better performance across types in Experiment 1 rather than the observed result of better performance in Experiment 2. Perhaps our results can be reconciled with this previous work in that these previous studies often give participants an extensive study phase where they are taught category membership ahead of the experimental phase, where participants attempt to recall category membership of relatively complex categories (e.g., matching a painting with the correct artist; Kang & Pashler, 2012; Kornell & Bjork, 2008). In the current work, participants are learning an underlying structure (and associated category rules) rather than memorizing individual exemplars of a given category. The ability to learn the correct diagnostic rules for category membership in the current study may provide an advantage in cases where only one category type is presented (Experiment 2). Participants may never master the membership rules for each category type when presented with mixed types (Experiment 1). However, this is only our best guess at

describing these discrepancies with previous research, and more work is needed to tease apart the complex dynamics of massed versus distributed practice and the relationship each has with learning over time.

It is helpful to consider how concept learning theories and their associated formal models account for the current results. Models that are able to (a) make predictions at the individual difference level and (b) make predictions about relative degree of discrimination over time are particularly useful in interpreting the current results. Our discussion in this section will focus on two particular models, the generalized invariance structure theory model (GISTM; Vigo, 2013, 2014) and the generalized context model (GCM; Nosofsky, 1984). We do this because we decided to focus on difficulty as a key construct and on models that give difficulty ratings for concepts as operationalized by either error rates or RTs. However, existing RT models like the exemplar-based random walk model (EBRW; Nosofsky & Palmeri, 1997; Palmeri, 1997) model purport to predict individual decisions and how these decisions change over time. For example, the EBRW does not give difficulty rating for concept structures, so we focused on models that do. Additionally, the context effects being studied are really best addressed with the GISTM and the GCM, for both are context models. Further, we do not believe it would be possible to apply the RT-specific models in this case because we did not give everyone the same category instances (recall that instances were pooled from all possible instances at random).

The GCM (Nosofsky, 1984) accurately predicts the $3_2[4]$ category learning difficulty ordering. The core underlying processes of similarity, choice, and attention are incorporated in Nosofsky's (1984) account through the use of ideas from multidimensional scaling (Shepard, 1962, 1974), choice probability theory (similarity choice model; Luce, 1963), and a previous exemplar-based account of category learning (Medin & Schaffer, 1978), respectively. Another major theoretical contribution of the GCM involves the use of attention parameters (one per stimulus dimension). Using 13 free parameter values (two attention parameter values per each of the six structures and one scaling parameter), the GCM accounts for 100% of the variance per session of Experiment 1 ($r^2 = 1$) and 100% of the data per session of Experiment 2 ($r^2 = 1$). This is not surprising because 13 parameters are used to predict six data points; thus, the model is likely overfitting the data. The estimated parameters of the model are given in Appendix 1. These were estimated using Solver, a parameter estimation program in Excel 2016. Presumably, via the use of the discrimination parameter c , the GCM suggests that increased discrimination to object exemplars within each category type is contributing to reduced errors and reduced response times over sessions. However, estimates for this parameter are nearly the same value for each of the sessions.

Thus, the GCM does not seem to capture the observed learning effect via its discrimination parameter c . Further, via attention weights per dimension, the model predicts that observers are paying more attention to diagnostic dimensions of the category over time and with practice. If this is indeed the case, then we would expect the estimated parameters to behave in a predictable manner across sessions of the current experiment. However, the attention weights do not behave consistently with the theory, as demonstrated by the tables in Appendix 1.

In contrast, generalized invariance structure theory, or GIST (Vigo, 2013, 2014), proposes an alternative notion of concept learning where observers are invariance pattern detectors. In other words, observers detect abstract symmetries inherent in the dimensional structure of a category of objects in order to determine the degree of diagnosticity and/or redundancy of each of its dimensions. The observer is then able to ascertain or assess which dimensions should be used in the formation of concept learning rules. As such, the ability to detect invariance patterns in categorical stimuli is a necessary precursor to concept formation in GIST. The core model of the theory is referred to as the “generalized invariance structure theory model,” or GISTM. Without free parameters, the model is able to account for about 90% (and above) of the variance in data from recent and historical categorization experiments. Discrimination in the GISTM refers to the ability to discriminate between invariance-based relational or structural information (represented as memory traces referred to a “ideotypes”) between categories rather than between object exemplars within categories, as in the GCM. Although there is a version of the GISTM without free parameters, it is not useful when analyzing individual differences in concept learning performance. Instead, we use the variant of the GISTM that incorporates a distance scaling parameter k (its only free parameter), which serves to account for individual differences regarding the ability to extract invariance patterns. That is, by indicating the overall degree of discrimination between the invariance of differing category structures, k serves as an index that summarizes classification performance. Technically, k indicates the overall degree of discriminability between the ideotypes and the standard ideotype represented by the zero point in the coordinates of psychological space. The standard ideotype represents category structures with zero degree of invariance. Larger k values indicate better discrimination and relatively higher discriminability between distances means relatively higher sensitivity to invariance patterns.

The ability of the scaling parameter k to reflect group differences is well demonstrated in recent work reporting differences in classification performance between populations. Vigo et al. (2015) tested college-aged adults, non-ADHD adolescents, and adolescents with ADHD on a classification task that necessitated learning geometric objects in a learning

phase and subsequently classified members and nonmembers. Of particular interest for our purposes are the estimated values of k for each group, which were reported as follows: adults ($k = .76$); non-ADHD adolescents ($k = .54$); adolescents with ADHD ($k = .29$). These k values might be interpreted to indicate relative degrees of overall discrimination, where higher values represent better overall discrimination.

We applied the GISTM-SE to the data from the current experiments (for further details, see Vigo, 2013, 2014). See Appendix 2 for the k parameter estimates and overall model fits. The parameter values were estimated using Solver, a parameter estimation program in Excel 2016. Note that with the use of only one parameter value per session (as opposed to 13 for the GCM), the model accounts for the variance in the data very well. Furthermore, and as shown, these parameter values are consistent with GIST in that they indicate a gradual increase in overall discrimination by participants across sessions and reflect the overall improvement in performance reported above. The parameter k encapsulates scaling, discrimination, and learning rate in the current study. The model suggests that the detection of invariance patterns is what observers are learning across sessions and discriminate between in the current learning tasks. In other words, observers are gradually becoming more sensitive to these invariance patterns over time.

Further, these results have implications for those interested in expertise development if one views expertise as the normative outcome of learning (see Palmeri & Cottrell, 2009). In this way, the development of expertise (in terms of its behavioral markers) can be described by accurately and precisely describing changes in knowledge representation that results from learning. Viewing expertise as the endpoint of learning affords the use of formal concept learning models as well as their associated representational theories as a theoretical base for the study of expertise development. The way in which classification performance unfolds over time due to experience and training is fundamental to research on both the acquisition of perceptual skills (i.e., perceptual expertise) and expert classification. For instance, the difference between novice and expert classification decisions have been examined in jobs that require years of studying highly specialized image sets, including radiology (Drew et al., 2013; Evans, Georgian-Smith, Tambouret, Birdwell, & Wolfe, 2013; Myles-Worsley et al., 1988; Evans et al., 2011), cytology (Evans, Wolfe, Tambouret, & Wilbur, 2010), and dermatology (Norman, Rosenthal, Brooks, Allen, & Muzzin, 1989). The reported relationship between classification accuracy and the time course of classification decisions throughout the learning process may have implications for training in these fields.

Limitations of the current experiments necessitate future replications and extensions in this area. For instance, the current work tested classification performance in a rather specific domain, using the $3_2[4]$ category structure types in a parainformative task that did not include corrective feedback. Advantages of using such a precise method include high internal validity and results that are easily comparable to many previous studies. However, it is likely that such a well-defined specific methodology also reduces the generalizability of the results. Accordingly, future research should employ several methodological extensions in order to expend the current findings. Future studies might use a seroinformative task (instead of a parainformative task) and/or provide corrective feedback to participants. Perhaps most critical, however, is that similar studies be conducted using category families that vary in their size and dimensionality by sampling from the 84 category structures studied by Vigo (2013). Additionally, this work should use stimuli that better resemble real-world categories. It should be noted that recent work has already begun to do so (Vigo & Doan, 2015; Vigo et al., 2015). The use of well-defined category structures is a scientifically responsible starting point since they represent the simplest cases and provide clear quantities that are comparable across experiments. However, we are assuming that the same cognitive and perceptual processes taking place during these experiments also take place in the real world when processing everyday stimuli. This is an assumption that should be tested in future work. In addition, although the choice of testing participants on three consecutive days is based on convention, it is likely an arbitrary interval. As such, future work might explore the influence of varying intervals between sessions of classification learning.

In conclusion, the two experiments above tested the stability of the learning difficulty ordering for the $3_2[4]$ category structure and explored the relationship between accuracy and response times when participants are tested over a period of extended learning. We found that when category structures are randomly sampled from across types there is stability in accuracy and response times across sessions. However, when participants are given instances of a single type, the ordering varies across sessions in terms of accuracy but can still be revealed by response times. Further, we found that the GISTM-SE was able to fit our data accurately and that GIST accounts for differences in learning and learning effects, in general. It is likely that many lines of empirical work in experience-based learning might be unified under this theoretical framework. It is our hope that this work inspires future research on concept learning and that the robustness of these empirical and theoretical relationships are tested in new domains of application.

Appendix 1 GCM Parameter Estimates

Errors					Response Times			
Experiment 1 – Session 1								
	<i>c</i>	w1	w2	w3	<i>c</i>	w1	w2	w3
I	23.96	0.32	0.35	0.33	24.10	0.67	0.00	0.33
II	23.96	0.33	0.33	0.33	24.10	0.33	0.33	0.33
III	23.96	0.28	0.39	0.33	24.10	0.28	0.39	0.33
IV	23.96	0.28	0.39	0.33	24.10	0.28	0.38	0.33
V	23.96	0.38	0.28	0.33	24.10	0.38	0.29	0.33
VI	23.96	0.33	0.33	0.33	24.10	0.33	0.33	0.33
Experiment 1 – Session 2								
I	23.94	0.31	0.36	0.33	23.96	0.36	0.31	0.33
II	23.94	0.33	0.33	0.33	23.96	0.33	0.33	0.33
III	23.94	0.28	0.38	0.33	23.96	0.28	0.38	0.33
IV	23.94	0.28	0.39	0.33	23.96	0.28	0.39	0.33
V	23.94	0.31	0.36	0.33	23.96	0.38	0.29	0.33
VI	23.94	0.33	0.33	0.33	23.96	0.33	0.33	0.33
Experiment 1 – Session 3								
I	23.94	0.31	0.35	0.33	23.96	0.33	0.33	0.33
II	23.94	0.33	0.33	0.33	23.96	0.33	0.33	0.33
III	23.94	0.28	0.38	0.33	23.96	0.28	0.39	0.33
IV	23.94	0.28	0.39	0.33	23.96	0.28	0.39	0.33
V	23.94	0.31	0.36	0.33	23.96	0.38	0.29	0.33
VI	23.94	0.33	0.33	0.33	23.96	0.33	0.33	0.33
Experiment 2 – Session 1								
I	23.96	0.31	0.36	0.33	24.10	0.67	0.00	0.33
II	23.96	0.33	0.33	0.33	24.10	0.33	0.33	0.33
III	23.96	0.28	0.39	0.33	24.10	0.30	0.37	0.33
IV	23.96	0.28	0.39	0.33	24.10	0.29	0.38	0.33
V	23.96	0.39	0.28	0.33	24.10	0.37	0.29	0.33
VI	23.96	0.33	0.33	0.33	24.10	0.33	0.33	0.33
Experiment 2 – Session 2								
I	23.92	0.31	0.36	0.33	23.96	0.35	0.32	0.33
II	23.92	0.33	0.33	0.33	23.96	0.33	0.33	0.33
III	23.92	0.27	0.40	0.33	23.96	0.29	0.38	0.33
IV	23.92	0.27	0.40	0.33	23.96	0.28	0.39	0.33
V	23.92	0.30	0.37	0.33	23.96	0.39	0.28	0.33
VI	23.92	0.33	0.33	0.33	23.96	0.33	0.33	0.33
Experiment 2 – Session 3								
I	23.93	0.31	0.35	0.33	23.96	0.35	0.32	0.33
II	23.93	0.33	0.33	0.33	23.96	0.33	0.33	0.33
III	23.93	0.27	0.39	0.33	23.96	0.28	0.39	0.33
IV	23.93	0.27	0.40	0.33	23.96	0.28	0.39	0.33
V	23.93	0.30	0.37	0.33	23.96	0.39	0.28	0.33
VI	23.93	0.33	0.33	0.33	23.96	0.33	0.33	0.33

Appendix 2 GISTM-SE Parameter Estimates

Experiment 1

	Errors	Response Times
	<i>k</i>	<i>k</i>
Session 1	1.23	1.01
Session 2	1.82	1.29
Session 3	2.03	1.56
	$R^2 = 0.914$	$R^2 = 0.849$

Experiment 2

	Errors	Response Times
	<i>k</i>	<i>k</i>
Session 1	0.81	1.11
Session 2	1.36	1.41
Session 3	1.74	1.56
	$R^2 = 0.794$	$R^2 = 0.831$

References

- Anderson, J. R. (1991). The adaptive nature of human categorization. *Psychological Review*, 98(3), 409.
- Bédard, J., & Chi, M. T. (1992). Expertise. *Current Directions in Psychological Science*, 1(4), 135–139.
- Bukach, C. M., Phillips, W. S., & Gauthier, I. (2010). Limits of generalization between categories and implications for theories of category specificity. *Attention, Perception, & Psychophysics*, 72(7), 1865–1874.
- Chi, M. T., Feltovich, P. J., & Glaser, R. (1981). Categorization and representation of physics problems by experts and novices. *Cognitive Science*, 5(2), 121–152.
- Cowan, N. (2001). Metatheory of storage capacity limits. *Behavioral and Brain Sciences*, 24(01), 154–176.
- Doan, C. A., & Vigo, R. (2016). Constructing and deconstructing concepts: On the nature of category modification and unsupervised sorting behavior. *Experimental Psychology*, 63(5), 249.
- Drew, T., Vo, M. L. H., Olwal, A., Jacobson, F., Seltzer, S. E., & Wolfe, J. M. (2013). Scanners and drillers: Characterizing expert visual search through volumetric images. *Journal of Vision*, 13(10), 3.
- Evans, K. K., Cohen, M. A., Tambouret, R., Horowitz, T., Kreindel, E., & Wolfe, J. M. (2011). Does visual expertise improve visual recognition memory?. *Attention, Perception, & Psychophysics*, 73(1), 30–35.
- Evans, K. K., Georgian-Smith, D., Tambouret, R., Birdwell, R. L., & Wolfe, J. M. (2013). The gist of the abnormal: Above-chance medical decision making in the blink of an eye. *Psychonomic Bulletin & Review*, 20(6), 1170–1175.
- Evans, K., Wolfe, J., Tambouret, R., & Wilbur, D. (2010). In the blink of an eye: Discrimination and localization of abnormalities in cervical cytology screening from a global signal. *Cancer Cytopathology*, 118(5), 389–389.
- Feldman, J. (2000). Minimization of Boolean complexity in human concept learning. *Nature*, 407(6804), 630–633.
- Feldman, J. (2003). A catalog of Boolean concepts. *Journal of Mathematical Psychology*, 47(1), 75–89.
- Gauthier, I., & Tarr, M. J. (1997). Becoming a “Greeble” expert: Exploring mechanisms for face recognition. *Vision Research*, 37(12), 1673–1682.
- Gluck, M. A., & Bower, G. H. (1988). Evaluating an adaptive network model of human learning. *Journal of Memory and Language*, 27(2), 166–195.
- Goodwin, G. P., & Johnson-Laird, P. N. (2011). Mental models of Boolean concepts. *Cognitive Psychology*, 63(1), 34–59.
- Hatala, R. M., Brooks, L. R., & Norman, G. R. (2003). Practice makes perfect: The critical role of mixed practice in the acquisition of ECG interpretation skills. *Advances in Health Sciences Education*, 8(1), 17–26.
- Kang, S. H., & Pashler, H. (2012). Learning painting styles: Spacing is advantageous when it promotes discriminative contrast. *Applied Cognitive Psychology*, 26(1), 97–103.
- Kemp, C. (2012). Exploring the conceptual universe. *Psychological Review*, 119(4), 685.
- Kornell, N., & Bjork, R. A. (2008). Learning concepts and categories: Is spacing the “enemy of induction”? *Psychological Science*, 19(6), 585–592.
- Kroger, J. K., Holyoak, K. J., & Hummel, J. E. (2004). Varieties of sameness: The impact of relational complexity on perceptual comparisons. *Cognitive Science*, 28(3), 335–358.
- Kruschke, J. K. (1992). ALCOVE: An exemplar-based connectionist model of category learning. *Psychological Review*, 99(1), 22.
- Kurtz, K. J., Levering, K. R., Stanton, R. D., Romero, J., & Morris, S. N. (2013). Human learning of elemental category structures: Revising the classic result of Shepard, Hovland, and Jenkins (1961). *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 39(2), 552.
- Lewandowsky, S. (2011). Working memory capacity and categorization: Individual differences and modeling. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 37(3), 720.
- Love, B. C. (2002). Comparing supervised and unsupervised category learning. *Psychonomic Bulletin & Review*, 9(4), 829–835.
- Love, B. C., & Markman, A. B. (2003). The nonindependence of stimulus properties in human category learning. *Memory & Cognition*, 31(5), 790–799.
- Love, B. C., Medin, D. L., & Gureckis, T. M. (2004). SUSTAIN: A network model of category learning. *Psychological Review*, 111(2), 309.
- Luce, R. D. (1963). A threshold theory for simple detection experiments. *Psychological Review*, 70(1), 61–79.
- Luck, S. J., & Vogel, E. K. (1997). The capacity of visual working memory for features and conjunctions. *Nature*, 390(6657), 279–281.
- Medin, D. L., & Schaffer, M. M. (1978). Context theory of classification learning. *Psychological Review*, 85(3), 207.
- Murphy, G. L. (2002). *The big book of concepts*. Cambridge: MIT Press.
- Myles-Worsley, M., Johnston, W. A., & Simons, M. A. (1988). The influence of expertise on X-ray image processing. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 14(3), 553.
- Norman, G. R., Rosenthal, D., Brooks, L. R., Allen, S. W., & Muzzin, L. J. (1989). The development of expertise in dermatology. *Archives of Dermatology*, 125(8), 1063–1068.
- Nosofsky, R. M. (1984). Choice, similarity, and the context theory of classification. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 10(1), 104.
- Nosofsky, R. M., Gluck, M. A., Palmeri, T. J., McKinley, S. C., & Gauthier, P. (1994). Comparing modes of rule-based classification learning: A replication and extension of Shepard, Hovland, and Jenkins (1961). *Memory & Cognition*, 22(3), 352–369.

- Nosofsky, R. M., Little, D. R., Donkin, C., & Fific, M. (2011). Short-term memory scanning viewed as exemplar-based categorization. *Psychological Review*, 118(2), 280.
- Nosofsky, R. M., & Palmeri, T. J. (1996). Learning to classify integral-dimension stimuli. *Psychonomic Bulletin & Review*, 3(2), 222–226.
- Nosofsky, R. M., & Palmeri, T. J. (1997). An exemplar-based random walk model of speeded classification. *Psychological Review*, 104(2), 266.
- Nosofsky, R. M., Palmeri, T. J., & McKinley, S. C. (1994). Rule-plus-exception model of classification learning. *Psychological Review*, 101(1), 53.
- 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(3), 608.
- Palmeri, T. J. (1997). Exemplar similarity and the development of automaticity. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 23(2), 324.
- Palmeri, T. J., & Cottrell, G. (2009). Modeling perceptual expertise. In I. Gauthier, M. Tarr, & D. Bub (Eds.), *Perceptual expertise: Bridging brain and behavior* (pp. 197–244). Oxford, UK: Oxford University Press.
- Rehder, B., & Hoffman, A. B. (2005). Eyetracking and selective attention in category learning. *Cognitive Psychology*, 51(1), 1–41.
- Rubenstein, A. (2013). Response time and decision making: An experimental study. *Judgment and Decision Making*, 8(5), 540.
- Scott, L. S., Tanaka, J. W., Sheinberg, D. L., & Curran, T. (2008). The role of category learning in the acquisition and retention of perceptual expertise: A behavioral and neurophysiological study. *Brain Research*, 1210, 204–215.
- Searston, R. A., & Tangen, J. M. (2017). Expertise with unfamiliar objects is flexible to changes in task but not changes in class. *PLoS ONE*, 12(6), e0178403.
- Shepard, R. N. (1962). The analysis of proximities: Multidimensional scaling with an unknown distance function. *Psychometrika*, 27(2), 125–140.
- Shepard, R. N. (1974). Representation of structure in similarity data: Problems and prospects. *Psychometrika*, 39(4), 373–421.
- Shepard, R. N., Hovland, C. I., & Jenkins, H. M. (1961). Learning and memorization of classifications. *Psychological Monographs: General and Applied*, 75(13), 1–42.
- Sims, V. K., & Mayer, R. E. (2002). Domain specificity of spatial expertise: The case of video game players. *Applied Cognitive Psychology*, 16(1), 97–115.
- Tanaka, J. W., Curran, T., & Sheinberg, D. L. (2005). The training and transfer of real-world perceptual expertise. *Psychological Science*, 16(2), 145–151.
- Tanaka, J. W., & Taylor, M. (1991). Object categories and expertise: Is the basic level in the eye of the beholder?. *Cognitive Psychology*, 23(3), 457–482.
- Vigo, R. (2006). A note on the complexity of Boolean concepts. *Journal of Mathematical Psychology*, 50(5), 501–510.
- Vigo, R. (2009). Categorical invariance and structural complexity in human concept learning. *Journal of Mathematical Psychology*, 53(4), 203–221.
- Vigo, R. (2011). Towards a law of invariance in human concept learning. In *Proceedings of the 33rd Annual Meeting of the Cognitive Science Society* (pp. 2580–2585). Austin, TX: Cognitive Science Society.
- Vigo, R. (2013). The gist of concepts. *Cognition*, 129(1), 138–162.
- Vigo, R. (2014). *Mathematical principles of human conceptual behavior: The structural nature of conceptual representation and processing* (Scientific Psychology Series). New York, NY: Routledge.
- Vigo, R., & Doan, C. A. (2015). The structure of choice. *Cognitive Systems Research*, 36, 1–14.
- Vigo, R., Doan, K. M. C., Doan, C. A., & Pinegar, S. (2018). On the learning difficulty of visual and auditory modal concepts: Evidence for a single processing system. *Cognitive Processing*, 19(1), 1–16.
- Vigo, R., Evans, S. W., & Owens, J. S. (2015). Categorization behaviour in adults, adolescents, and attention-deficit/hyperactivity disorder adolescents: A comparative investigation. *The Quarterly Journal of Experimental Psychology*, 68(6), 1058–1072.
- Vigo, R., & Zeigler, D. E. (2016). On the relationship between perceived structural complexity and temporal judgments. In J. Houpt & L. Blaha (Eds.), *Mathematical models of perception and cognition, Volume 2: A Festschrift for James T. Townsend* (pp. 47–71). New York, NY: Psychology Press.
- Vigo, R., Zeigler, D. E., & Halsey, P. A. (2013). Gaze and informativeness during category learning: Evidence for an inverse relation. *Visual Cognition*, 21(4), 446–476.
- Voss, J. F., Greene, T. R., Post, T. A., & Penner, B. C. (1983). Problem-solving skill in the social sciences. *The Psychology of Learning and Motivation*, 17, 165–213.