Deficits in Category Learning in Older Adults: Rule-Based Versus Clustering Accounts

Stephen P. Badham

Nottingham Trent University, UK

Adam N. Sanborn & Elizabeth A. Maylor

University of Warwick, UK

Author Note

Stephen P. Badham, Department of Psychology, Nottingham Trent University, United Kingdom; Adam N. Sanborn and Elizabeth A. Maylor, Department of Psychology, University of Warwick, Coventry, United Kingdom.

This research was supported by a grant awarded to SPB and EAM from the UK's Economic and Social Research Council (ESRC; ES/K002732/1) and a grant awarded to ANS from the ESRC (ES/K004948/1). We thank Wolf Vanpaemel for helpful comments on an earlier version of the manuscript. The complete dataset and model reported here can be found at http://doi.org/10.17605/OSF.IO/M7TCK.

Correspondence concerning this article should be addressed to Stephen P. Badham, College of Business Law & Social Sciences, School of Social Sciences, Department of Psychology, Nottingham Trent University, Burton Street, Nottingham, NG1 4BU, United Kingdom. E-mail: stephen.badham@ntu.ac.uk

Abstract

Memory research has long been one of the key areas of investigation for cognitive aging researchers but only in the last decade or so has categorization been used to understand age differences in cognition. Categorization tasks focus more heavily on the grouping and organization of items in memory, and often on the process of learning relationships through trial and error. Categorization studies allow researchers to more accurately characterize age differences in cognition: whether older adults show declines in the way in which they represent categories with simple rules or declines in representing categories by similarity to past examples. In the current study, young and older adults participated in a set of classic category learning problems, which allowed us to distinguish between three hypotheses: (i) rule-complexity: categories were represented exclusively with rules and older adults had differential difficulty when more complex rules were required, (ii) rule-specific: categories could be represented either by rules or by similarity, and there were age deficits in using rules, and (iii) clustering: similarity was mainly used and older adults constructed a less-detailed representation by lumping more items into fewer clusters. The ordinal levels of performance across different conditions argued against rule-complexity, as older adults showed greater deficits on less complex categories. The data also provided evidence against rule-specificity, as single-dimensional rules could not explain age declines. Instead, computational modelling of the data indicated that older adults utilized fewer conceptual clusters of items in memory than did young adults.

Keywords: aging, categorization, learning, task complexity, clustering

Deficits in Category Learning in Older Adults: Rule-Based Versus Clustering Accounts

Categorization is the process of grouping and organizing sensory information and draws upon many constructs in cognitive science including learning, decision making, reasoning and attention (Pothos & Wills, 2011). Understanding how individuals form categories from patterns in the environment is central to human learning (Feldman, 2000) and is relevant to a variety of circumstances in everyday life: Is it high or low fat? Are their policies left or right wing? Will this medication raise or lower blood pressure?

Surprisingly, given the extensive research into age differences in memory (e.g., Naveh-Benjamin & Ohta, 2012), there has been far less research into how young and older adults differ in the learning of categorical information (cf. Maintenant, Blaye, & Paour, 2011). Categorization research has the potential to deliver new insight into age-related changes in memory because these tasks can involve precise manipulations of the structure of categories in order to reveal the underlying representation. Therefore, categorization tasks can better assess the details of learning and the interference between competing items in memory than can most memory tasks.

A main point of contention in category learning is whether individuals are using rules or similarity in order to make their judgments. Rule-based approaches classically assume that there is a set of features that describe a category and a new stimulus is either entirely a category member or not (Bourne, 1970; Bruner, Goodnow, & Austin, 1986; Nosofsky, Palmeri, & McKinley, 1994). In contrast, similarity-based approaches assume that a new stimulus is compared either directly to exemplars experienced in the past, or to a single prototype of these past examples, producing a graded category membership (Medin & Schaffer, 1978; Nosofsky, 1986; Reed, 1972). The most flexible similarity-based approach is clustering, which because it clusters past exemplars into multiple prototypes, can produce representations that match exemplar models, prototype models, or anywhere in between (Anderson, 1991; Love, Medin, & Gureckis, 2004; Rosseel, 2002; Vanpaemel & Storms, 2008).

Initial conceptions of categorization were rule-based, but following theoretical and empirical arguments for graded category membership, similarity approaches became standard (Rosch, 1973; Wittgenstein, 1953). Later research leveraged the strengths of both rule-based and similarity-based categorization, through the development of hybrid models that have both an explicit rule-based system and an implicit similarity-based system (Ashby, Alfonso-Reese, Turken, & Waldron, 1998). Although there are empirical effects in category learning that point to both rule-based and similarity-based representations, it is possible that deficits in category learning in older adults are just of one type. Therefore, our question is: are age deficits best described as deficits in rule-based categorization or as deficits in similarity-based categorization?

Investigations into categorization deficits in older adults have compared young and older adults across various category structures to determine where the deficits for older adults lie, exploring both rule-based and clustering accounts. One rule-inspired hypothesis is that older adults are differentially worse at more complex categories (Cerella, Poon, & Williams, 1980), which we will term **rule-complexity**. For example, in Racine, Barch, Braver, and Noelle (2006) one category was composed of examples lying at the extremes of the space of possible continuous-feature stimuli, while the other was composed of examples lying in the middle of the space of possible stimuli. Participants were told what rule to follow, where category membership was defined by either a two- (low complexity) or three- (high complexity) part conjunctive rule. Racine et al. found that older adults performed differentially worse on the categories defined by more complex rules. Other categorization studies have supported the rule-complexity hypothesis by showing that as the task becomes more difficult, age-related deficits increase (so long as floor/ceiling effects are avoided). Additionally, studies involving functional relations (which are similar to categorization tasks in that participants must learn rules linking stimuli to responses) demonstrate greater age-related deficits for more complex relations, such as inverse (Griego & Kliegel, 2007) and multiplicative (Chasseigne & Lafon, 2002) relations.

A different rule-inspired hypothesis for age deficits follows from the model COVIS (Ashby et al., 1998). COVIS is a hybrid model consisting of two systems: an explicit system that can learn simple rules and an implicit system that can be considered similarity-based. It has been argued that categorization deficits in older adults (Rabi & Minda, 2016) and children (Minda, Desroches, & Church, 2008), relative to young adults, are larger for complex rule-based categorization tasks compared to implicit categorization tasks. Likewise, when increasing the number of irrelevant dimensions in a rule-based categorization task, Filoteo, Maddox, Ing, Zizak, and Song (2005) found a trend for older participants to perform differentially worse. Older adults may therefore have difficulty with explicit, rule-based categories which are arguably more reliant on effortful processing. Age-related memory deficits are generally reduced or absent for implicit tests of memory (La Voie & Light, 1994; Light, Prull, La Voie, & Healy, 2000), where effortful strategic encoding and retrieval processes are not required. Furthermore, age deficits in executive prefrontal processing (West, 1996) of rules have been used to describe older adults’ poor performance at the Wisconsin Card Sorting Test (Rhodes, 2004). Therefore, it seems that a dual-system account such as COVIS could explain age deficits in categorization in only its rule-based system but not the implicit system, a hypothesis we term **rule-specificity**. However, other researchers have shown the contrary effect: a larger age deficit in the implicit system than in the rule-based system (e.g., Filoteo & Maddox, 2004; Mata, von Helverson, Karlsson, & Cüpper, 2012).

In contrast to these rule-based accounts is the possibility that age deficits are implicit, and particularly the notion that older adults may not generate as detailed an implicit category representation as do young adults (Love & Gureckis, 2007), a hypothesis we call **clustering**. The assumption behind this hypothesis is that people use multiple prototypes (e.g., clusters) to represent categories, and the more clusters that are used the more detailed the category representation can be. Studies have shown that older adults can construct simple prototype representations as well as can young adults, but do not represent complex categories with as much detail as do young adults (Hess, 1982; Hess & Slaughter, 1986). Also, older adults have poorer memory for category members that are exceptions to rules (Davis, Love, & Maddox, 2012; Love & Gureckis, 2007). For example, Davis et al. (2012) showed participants images of beetles that were categorized into two groups. The features of the beetles were arranged such that the majority of beetles in one group would possess a given feature (e.g., thick legs) but a small subset would have the opposite feature (e.g., thin legs – an exception to the rule). Older adults showed a deficit relative to young adults when categorizing these exception stimuli. This can be explained as older adults constructing fewer clusters than young adults to represent categories.

In summary, we have identified three hypotheses related to age differences in categorization: (i) rule-complexity: differential difficulty with category structures defined by more complex rules in a rule-only categorization model, (ii) rule-specific: age deficits in the use of explicit rule-based but not implicit systems of a hybrid model, and (iii) clustering: a tendency for older adults to construct fewer clusters in similarity-based categories. These explanations are difficult to tease apart: they can imitate one another quite closely as more complex categories also generally require both more complex rules and more clusters in order to be represented accurately. Researchers have only begun to test these accounts against one another: Rabi and Minda (2016) compared the two rule-based accounts and found evidence to support rule-specificity over rule-complexity. The key evidence was smaller age deficits in a more complex categorization task compared to a less complex categorization task. However, this study did not rule out age deficits due to clustering. The current study aimed to replicate and further explore this key empirical finding to determine if a rule-specificity account of deficits is plausible, and also to establish if the empirical age deficits could be explained better with a clustering account.

The Current Study

Here, we compare the rule-complexity, rule-specificity, and clustering hypotheses of age deficits against one another using a seminal paradigm from the categorization literature, the category learning problems of Shepard, Hovland, and Jenkins (1961). In this task, participants learn to place a series of eight geometric images into two categories across a series of learning blocks. The eight images were formed by factorial combinations of three binary dimensions (see top panel of Figure 1), which were form (square/triangle), color (black/white) and size (large/small). Four of the shapes were assigned to an ‘alpha’ group and four to a ‘beta’ group.

Shepard et al. (1961) identified six meaningfully distinct ways to form two groups of four stimuli from the set of eight geometric images (Types I, II, III, IV, V and VI). These groupings are based upon categorization rules of varying complexity and Types I to IV were used in the current study (see bottom panel of Figure 1). Type I is the simplest condition where a single dimension defines category membership (e.g., all the black images are in the alpha category and all the white images are in the beta category) and the other dimensions (e.g., size and form) are irrelevant. Type II defines category membership by two dimensions (e.g., black triangles and white squares are in the alpha group) with one irrelevant dimension (e.g., size). Type III uses all three dimensions to define category membership and categories are defined by a rule with an exception (e.g., all the black objects are in the alpha group apart from the small black square). Type IV also uses all three dimensions and all category members share the majority of their features with other category members (e.g., most of the large, black and triangular shapes are in the alpha group). Types III and IV seem similar and indeed often lead to similar levels of performance (e.g., Shepard et al., 1961) but one key difference is that participants can respond with 75% accuracy by paying attention to *any* single dimension for Type IV but can only achieve 75% accuracy in Type III with a single dimension for two out of the three dimensions (e.g., responding on the basis of color or form alone for Type III in Figure 1 would yield 75% accuracy but size would yield 50% accuracy).

In young adults, performance generally decreases from Type I to Type IV (Type I > Type II > Type III = Type IV; Kurtz, Levering, Stanton, Romero, & Moris, 2013; Nosofsky, Gluck, Palmeri, McKinley, & Glauthier, 1994; Shepard et al., 1961). For the rule-complexity hypothesis (Cerella et al., 1980), the prediction is simply that – unless there are floor/ceiling effects – age differences will follow this same pattern, that is, increasing age differences from Type I to Type IV. This hypothesis about rule-complexity based on learning difficulty is bolstered by formal mathematical analyses of the complexity of the rules needed to learn Types I-IV. Feldman (2000) introduced an explicit Boolean complexity measure of the Shepard et al. (1961) types, finding that this formal measure of complexity corresponded fairly closely to learning difficulty. Although there is some disagreement about the relative difficulty of Type III, Boolean complexity and various other measures of complexity agree that Type IV is more complex than Type II which itself is more complex than Type I (Goodman, Tenenbaum, Feldman, & Griffiths, 2008; Vigo, 2006; 2009). Therefore, both mathematical and behavioral accounts of complexity would predict greater age differences for Type IV than for Type II and greater age differences for Type II than for Type I.

Rabi and Minda (2016) found age differences that clearly went against the predictions of rule-complexity. Whereas young adults showed better performance on Type II than Type IV, older adults showed the opposite: their Type IV performance exceeded their Type II performance. The deficit in Type II was taken as evidence of rule-specificity in age deficits. It was argued that older adults were generally not able to use multi-dimensional rules, because of their poor overall performance on Type II. Also it was argued that older adults were unable to transition to the more flexible implicit system, and so "applied single-dimensional rules during Type II learning, but frequently switched rules during the course of the task to avoid negative feedback" (p. 194). We thus formulate the rule-specificity hypothesis to mean that older adults cannot use multi-dimensional rules, and must use either single-dimensional rules or their intact implicit system instead. This rule-specificity account was bolstered by an association between backward digit span and Type II performance, plausibly tying complex rule-based categorization to working memory capacity. Rule-specificity was also suggested to explain the reliable deficit older adults showed in Type IV performance: this was potentially a result of older adults following simple rules in Type IV rather than switching over to the more flexible implicit similarity-based system, as young adults do (e.g., Maddox et al., 2010).

Explanations based on deficits in rule use, however, are not the only kind of explanation for these age deficits. A different hypothesis was investigated by Davis et al. (2012) who found that older participants struggled much more with learning the exceptions to rules (summarized earlier). They proposed a clustering account of their results, that is, items were grouped into clusters, each of which is represented as a prototype of the items in a cluster. Clustering explanations essentially represent a category by multiple prototypes, which interpolates between the extremes of single prototype models and exemplar models (which represent all of the previously experienced items individually). Davis et al. supposed that older adults have more difficulty in constructing new clusters of items in memory, meaning that categories are represented more coarsely by older adults (see also Love & Gureckis, 2007). Their argument was bolstered by fitting their data with a clustering model of categorization, the Rational Model of Categorization (RMC; Anderson, 1991), and showing that the parameters indicated that older adults did not construct as many clusters as did young adults.

Intuitively, a clustering account could also explain the pattern of age deficits found by Rabi and Minda (2016). Types I and IV can both be represented well by a single cluster or prototype per category because the two categories in these tasks are linearly separable: a straight plane can be placed in the space of stimuli in Figure 1 for these two tasks that perfectly separates the two categories. In contrast, representing each category in Type II with a single cluster would be a catastrophe: because of the symmetric arrangement of the stimuli in each category, the prototype of each cluster would be exactly in the middle of the cube of stimuli, and so the inferred categories are indistinguishable and performance would be at chance. Thus the number of clusters is more critical in Type II than in Type I or Type IV, and so if older adults have greater difficulty constructing more clusters, then larger age differences are expected in Type II compared to Types I or IV, matching the empirical results.

Although it is intuitive that a clustering account can explain age deficits, we cannot know whether rule-specificity or clustering deficits better match human behavior until we evaluate them against data. We collected our own data in the Shepard et al. (1961) tasks, including Type III in addition to Types I, II, and IV, which first allowed us to determine if the pattern of age deficits replicated. Type III provides another benchmark against which to evaluate Types I, II, and IV, and an opportunity to see if Types III and IV are also equally difficult for older adults, as seen in young adults and as many complexity approaches predict.

Using these new data, we then evaluated the plausibility of the idea that older adults were using single-dimensional rules, employing a variety of measures. We finally fit the RMC to the trial-by-trial data to see if a clustering account could quantitatively match the data.

Method

Design

Young and older adults learned to categorize eight shapes into two groups. Each participant completed four conditions (Types I to IV) where group membership was determined by separate rules as outlined in the introduction.

Participants

Forty-eight young adults (42 female) aged 18–21 years (*M* = 19.3, *SD* = 0.7) and 48 healthy older adults (32 female) aged 60–87 years (*M* = 74.7, *SD* = 5.6) took part in the experiment. Ten of the older adults were in their 60s, 30 in their 70s, and eight in their 80s, with all except four aged 66-83. Young participants were recruited from the University of Warwick and received course credit. Older participants were active members of our Age Study Panel who were visited in their own homes and received £5 ($7); their self-rated eyesight, hearing, and general health averaged 4.1, 4.0, and 4.0 (equivalent to “good”), respectively, on a five-point scale (1 = “very poor” to 5 = “very good”). Participants were recruited in two batches (though all were tested within a seven-week period, January-March 2015), and the statistical implications of this are discussed in the results. All participants provided written informed consent, and the study was approved by the University of Warwick’s Humanities and Social Sciences Research Ethics Committee.

Young and older participants did not show evidence of a difference in their years of education, $t(53.87) = 1.88$, $BF_{10} = 1.01$[1] ($M_{\text{young}} = 14.01$, $SD_{\text{young}} = 0.93$; $M_{\text{older}} = 14.97$, $SD_{\text{older}} = 3.41$). To assess cognitive functioning, participants completed the Digit Symbol Substitution test from the Wechsler Adult Intelligence Scale – Revised (Wechsler, 1981) as a measure of processing speed, and the multiple choice part of the Mill Hill vocabulary test (Raven, Raven, & Court, 1988) as a measure of crystallized intelligence. The results were consistent with the literature (e.g., Salthouse, 2010): young adults performed better than older adults at the speed task, $t(94) = 10.98$, $BF_{10} = 2.42 \text{ x } 10^{15}$ ($M_{\text{young}} = 74.63$, $SD_{\text{young}} = 10.04$; $M_{\text{older}} = 51.96$, $SD_{\text{older}} = 10.19$), and older adults performed better than young adults at the vocabulary task, $t(94) = 8.99$, $BF_{10} = 1.95 \text{ x } 10^{11}$ ($M_{\text{young}} = 16.83$, $SD_{\text{young}} = 3.47$; $M_{\text{older}} = 23.67$, $SD_{\text{older}} = 3.96$).

Materials

Images of eight geometric shapes were constructed for use in the experiment. Large images had a base of width 250 pixels and small images had a base of width 125 pixels, corresponding to widths of approximately eight and four degrees of viewing angle on screen, respectively. Triangles were equilateral and both square and triangle image bases were horizontal. Images were presented in black or white on a mid-gray background.

Counterbalancing. The four conditions were within participants so this resulted in 24 possible test orders for Types I to IV. Additionally, each condition had several permutations (e.g., Type I had three permutations because category membership could be defined by color, form or size). Types II, III and IV had 3, 12 and 4 permutations, respectively. Twenty-four versions of the experiment were created (one for each test order) and the permutations of each type were randomly assigned to each version such that each permutation was used

[1] See later for explanation of the use of Bayes factors for comparisons.

equally across the experiment (for simplification, Type III was reduced to 3 permutations by assigning all but one dimension randomly). Category memberships ‘alpha’ and ‘beta’ were also randomly determined. These 24 versions of the experiment were then used four times (twice with young and twice with older adults).

Procedure

Participants were initially shown rule-based instructions taken verbatim from Kurtz et al. (2013) who found that such instructions are more likely to yield the typical Type II advantage (relative to Types III and IV) shown in the literature. These instructions encourage participants to ‘learn a rule that allows [them] to tell whether each example belongs in the alpha or beta category’ (Kurtz et al., 2013, p. 6). Participants were then shown a single screen containing all eight shapes (in no particular arrangement, and without any category information) so that they could clearly see the differences between the shapes. They were informed that these were all of the shapes that would be used in the experiment. Following this, they commenced the first condition of the experiment.

In each trial, an image was presented centrally on the screen. Participants were initially required to guess if it belonged in the alpha or beta category by pressing the keys “F” and “J” on the computer keyboard, which were relabelled ‘Alpha’ and ‘Beta’, respectively (the words Alpha and Beta were also displayed in the bottom left and right corners of the screen, respectively). The image remained on screen until a response was made, then after 500 ms of blank screen, feedback was provided. The image reappeared on screen and either ‘Correct!’ appeared above it in green or ‘Incorrect!’ in red. For both feedback options, below the feedback image appeared the correct response in blue, for example, ‘Answer = Alpha’. The feedback remained on screen until the participant pressed the spacebar, then a further 500 ms of blank screen was displayed before the next trial.

In the first two blocks, all eight shapes were presented in the first half of the block and then again in the second half. This limited the possibility of the same shape appearing in adjacent trials. In subsequent blocks, the eight shapes were presented twice in each block of 16 trials without any constraints. This ordering replicates the original Shepard et al. (1961) study. Participants completed the task for six blocks (96 trials) or until they reached a criterion of perfect performance in two consecutive blocks. Once a condition was complete, a message on screen indicated that ‘a new rule [would] determine which images belong to each category’. Participants could rest between conditions as they wished. The experiment continued until the participant had completed all four categorization conditions.

Results

During our data collection process, we found interesting trends (qualitatively identical to those we report below) after testing 48 participants (i.e., 24 young and 24 older), but the key comparison (namely, the age by condition interaction) did not reach the standard value for statistical significance. Therefore, we tested an additional 48 participants and stopped our experiment at that point. This stopping rule invalidates the *p*-values calculated using standard null hypothesis significance (e.g., Wagenmakers, 2007), so we report test statistics and effect sizes without the *p*-values. Instead we report Bayes factors, which provide a valid measure of the evidence provided by the data even when the rule for stopping data collection depends on the results of a test (Rouder, 2014). This measure even provides strong guarantees about how much an experimenter can influence the statistical results, in particular when finding evidence that favors the alternative hypothesis (Sanborn & Hills, 2014).

Standard null hypothesis significance tests assess the probability of a test statistic arising from the null hypothesis, limiting researchers to only evaluating the plausibility of the null, and leaving them in an awkward position if there is not enough evidence to reject the null. In contrast, Bayesian methods explicitly compare the probability of the null and alternative hypotheses on even ground, so that evidence can be found in favor of the alternative hypothesis, the null hypothesis, or neither (Gallistel, 2009; Rouder, Speckman, Sun, Morey, & Iverson, 2009). A common Bayesian measure of evidence is the Bayes factor (BF_{10}; Kass & Raftery, 1995) which provides an odds ratio for the alternative/null hypotheses (values < 1 favor the null hypothesis and values > 1 favor the alternative hypothesis). For example, a BF_{10} of 2.5 would indicate that the alternative hypothesis is 2.5 times more likely than the null and a BF_{10} of 0.40 would indicate the converse (see Jarosz & Wiley, 2014). Associating labels with these values is arbitrary, but in past work labels such as 'substantial', 'strong', and 'decisive' have been associated with Bayes factors of 3, 10, and 100, respectively (Wetzels et al., 2011). These Bayes factors were calculated using the JASP computer software (Love et al., 2015). All *t*-tests are two-tailed using the standard Cauchy prior width of 0.707. The Bayesian ANOVAs construct a model for each of the possible combinations of terms and we report $\mathrm{BF}_{\mathrm{inclusion}}$ for each term because it gives a summary of the evidence for including that term in the models.

For the accuracy data, where a condition was terminated early due to a participant reaching criterion, 100% accuracy was assumed for all subsequent uncompleted blocks (as is typical with this paradigm: e.g., Kurtz et al., 2013; Nosofsky, Gluck et al., 1994). Figure 2 shows the overall means for Blocks 1-6, while Figure 3 shows overall age differences, for Types I-IV. Performance accuracy was entered into a 2 (Age: young, older) x 4 (Condition: Types I to IV) x 6 (Block: 1-6) repeated measures ANOVA. Young adults were more accurate than older adults, $F(1, 94) = 48.86$, $MSE = 0.17$, $\eta_p^2 = .34$, $BF_{10} = 3.16 \times 10^{12}$. There was a main effect of condition,[2] $F(2.59, 243.10) = 129.83$, $MSE = 0.07$, $\eta_p^2 = .58$, $BF_{10} > 10^{12}$, with Type I learned better than all other conditions (performance on Types II to IV is

[2] Throughout this article, any violations of sphericity in ANOVAs were corrected using the Greenhouse-Geisser correction.

investigated further below). A main effect of block showed that performance improved over time, $F(3.31, 311.55) = 138.91$, $MSE = 0.02$, $\eta_p{}^2 = .60$, $BF_{10} > 10^{12}$. Age interacted with condition, $F(2.59, 243.1) = 3.26$, $MSE = 0.07$, $\eta_p{}^2 = .03$, $BF_{10} = 7651$, but there was no evidence that it interacted with block, $F(3.31, 311.55) = 4.02$, $MSE = 0.02$, $\eta_p{}^2 = .04$, $BF_{10} = 0.88$. There was sizeable evidence against the three-way interaction between age, condition and block, $F(10.64, 999.82) = 3.31$, $MSE = 0.02$, $\eta_p{}^2 = .03$, $BF_{10} = 4 \text{ x } 10^{-3}$. As can be seen in Figure 2, the young adults' Type I performance was near ceiling, which could potentially be driving the age by condition interaction (see Figure 3).

In order to investigate potential age interactions without ceiling performance, the above ANOVA was repeated but with Type I excluded from the condition factor. Young adults performed better than older adults, $F(1, 94) = 53.78$, $MSE = 0.14$, $\eta_p{}^2 = .36$, $BF_{10} = 1.33 \text{ x } 10^{12}$, there was evidence for a main effect of condition, $F < 1$, $BF_{10} = 3.62$, and accuracy improved across blocks, $F(3.46, 324.82) = 87.48$, $MSE = 0.03$, $\eta_p{}^2 = .48$, $BF_{10} > 10^{12}$. There was an age by block interaction, $F(3.46, 324.82) = 7.61$, $MSE = 0.03$, $\eta_p{}^2 = .08$, $BF_{10} = 1.02 \text{ x } 10^4$, due to slower learning in older adults. Importantly, the age by condition interaction remained, $F(1.70, 159.94) = 3.61$, $MSE = 0.06$, $\eta_p{}^2 = .04$, $BF_{10} = 20.50$, confirming the different age-related deficits between Types II-IV evident in Figure 3; this interaction is investigated further below. There was evidence against the other interactions in the analysis (Condition x Block, $F < 1.36$, $BF_{10} = 2 \text{ x } 10^{-3}$, Age x Condition x Block, $F < 1$, $BF_{10} = 2.65 \text{ x } 10^{-5}$).

To interpret the above age by condition interaction, the condition by block (3 x 6) ANOVA was run separately for young and older adults. Older adults had a main effect of condition, $F(1.78, 83.51) = 4.83$, $MSE = 0.04$, $\eta_p{}^2 = .09$, $BF_{10} = 186.6$, but there was evidence that young adults did not, $F < 1$, $BF_{10} = 7.5 \text{ x } 10^{-2}$. T-tests (collapsed across blocks) revealed that older adults performed best at Type IV, $M = 0.62$, $SD = 0.08$, which was better than Type

II, *M* = 0.57, *SD* = 0.11, *t*(47) = 3.10, BF_{10} = 10.04, and possibly better than Type III, *M* = 0.59, *SD* = 0.08, *t*(47) = 2.36, BF_{10} = 1.92 (whereas Type II and III performance appeared the same, *t* < 1, BF_{10} = 0.24). Numerically, young adults performed best at Type II, *M* = 0.74, *SD* = 0.16, but there was evidence that performance did not differ from that in Type III, *M* = 0.72, *SD* = 0.12, *t*(47) < 1, BF_{10} = 0.229, and Type IV, *M* = 0.72, *SD* = 0.11 *t*(47) < 1, BF_{10} = 0.232. There was also evidence that performance did not differ between Type III and Type IV, *t*(47) < 1 , BF_{10} = 0.157.

Testing for Rule Use

The rule-specificity hypothesis is that older adults show deficits in the rule-based system, but have an intact implicit system. Because older adults perform worse across all four types, the rule specificity hypothesis implies that all of these declines are due to worse rule-based categorization. In particular, Rabi and Minda (2016) hypothesized that older adults are only rarely able to use conjunctive or disjunctive rules and instead must rely on single-dimensional rules. This can explain the superior performance that older adults demonstrated on Type IV versus Type II: any single-dimensional rule would result in 75% accuracy for Type IV, but result in 50% accuracy for Type II.

We first investigated whether conjunctive and disjunctive, or single-dimensional rules were used by looking at the consistency with which individuals were adhering to these rules in each of the four problems. To do so, we created a measure that is diagnostic as to whether single-dimensional rules are being used. First, we computed the number of mismatches (i.e., Hamming distance) between the responses in each block and the responses that would have been made using each of the three possible single-dimensional rules. Then the minimum of the three Hamming distances in each block was taken as the measure of adherence to the closest single-dimensional rule. The result is a score for each individual in each block, and the mean scores for the two age groups over the blocks are shown in Figure 4. Here, perfect performance would result in (minimum) Hamming distances of zero for Type I, eight for Type II, and four for both Types III and IV. If participants are consistently using a single-dimensional rule for any problem, then the Hamming distance will be zero.

For each type, a 2 (Age: young, older) x 6 (Block: 1-6) repeated measures ANOVA was conducted. For Type I, older adults had larger Hamming distances than young adults, $F(1, 94) = 7.57$, $MSE = 9.36$, $\eta_p{}^2 = .08$, $BF_{10} = 4.55$, and the Hamming distances decreased across blocks showing a trajectory towards the correct distance of zero, $F(3.10, 291.20) = 81.31$, $MSE = 1.52$, $\eta_p{}^2 = .46$, $BF_{10} > 10^{12}$, with no interaction, $F(3.10, 291.20) = 1.75$, $MSE = 9.36$, $\eta_p{}^2 = .02$, $BF_{10} = 0.534$.

For Type II, older adults had smaller Hamming distances than young adults, $F(1, 94) = 21.02$, $MSE = 5.63$, $\eta_p{}^2 = .18$, $BF_{10} = 1.42 \text{ x } 10^7$, the Hamming distances increased across blocks, $F(4.31, 405.15) = 15.30$, $MSE = 1.60$, $\eta_p{}^2 = .14$, $BF_{10} > 10^{12}$, and to a greater extent in young compared with older adults, $F(4.31, 405.15) = 10.11$, $MSE = 1.60$, $\eta_p{}^2 = .10$, $BF_{10} = 2.79 \text{ x } 10^9$. Interestingly, post hoc tests revealed that young adults showed a trajectory towards the correct distance of eight across blocks, $F(3.37, 158.49) = 25.16$, $MSE = 1.97$, $\eta_p{}^2 = .35$, $BF_{10} = > 10^{12}$, but there was evidence that older adults remained constant across blocks, $F(5, 235) = 1.17$, $MSE = 1.43$, $\eta_p{}^2 = .02$, $BF_{10} = 0.05$. Thus, it appears that older adults were neither trending toward using single-dimensional rules consistently, nor trending toward using the correct multi-dimensional rules consistently. Their responses in Type II problems were stuck between these two extremes, and did not change across blocks.

For Type III, there was no effect of age, $F < 1$, $BF_{10} = 0.153$, the Hamming distances did not change across blocks, $F < 1$, $BF_{10} = 0.003$, and there was no interaction, $F(5, 470) = 1.95$, $MSE = 1.47$, $\eta_p{}^2 = .02$, $BF_{10} = 6.93 \text{ x } 10^{-4}$.

For Type IV, there was no effect of age, F = 1.00, BF_{10} = 0.161, the Hamming distances decreased across blocks, F(5, 470) = 3.52, MSE = 1.44, η_p^2= .04, BF_{10} = 1.693, and there was no interaction, F < 1, BF_{10} = 0.018.

For Types III and IV, all participants were close to the Hamming distance that perfect performance would produce across all blocks. However, this only shows that their responses showed the right amount of deviation from single-dimensional rules – clearly the actual responding of both young and older adults was far from perfect for these two types (see Figure 2).

From the Hamming distance measures, older adults appear to be unable to learn the multi-dimensional rules required for Type II problems, and also do not appear to be using single-dimensional rules consistently instead. Of course, older adults may not be using single-dimensional rules consistently throughout a block as the Hamming distance measures, but instead are quickly switching between single-dimensional rules as they accumulate negative feedback (Ashby et al., 1998; Rabi & Minda, 2016). Fortunately, the Shepard et al. (1961) stimuli allow us to assess how often quick switches in single-dimensional rules are occurring by looking for consecutive trials in which the stimuli are maximally distant from one another (i.e., in Figure 1 pairs of stimuli that are in the opposite corners of the cube from one another). Looking at these consecutive trials (which make up 13% of all trials), participants who use the same single-dimensional rule in the two trials will always make two different responses, no matter which single-dimensional rule is used.

Figure 5 shows the proportion of trials on which young and older participants made the different responses to maximally different stimuli on consecutive trials, meaning that the two responses were consistent with using the same single-dimensional rule. The other types are included for completeness, but Type II is the most interesting task in this analysis because of the possibility that older adults are quickly switching between single-dimensional rules as they are unable to use multiple dimensional rules. In Type II only there is also a clear contrast between correct responding and consistent single-dimensional rule use: correct responding predicts a value near 0 while single-dimensional rule use predicts a value near 1. As shown in Figure 5, young adults make the same response more than half the time, while older adults make the same response almost exactly half the time (and only two older adults never made this response). Such a low percentage of different responses cannot result from consistent use of single-dimensional rules even across two consecutive trials; instead it looks most like randomly selecting a single-dimensional rule on each trial. What is particularly striking is that the proportion of different responses (indicating single-dimensional rule use) is only half, even when older adults made the correct response to the previous trial. This is notable because the COVIS explicit system, which is used as the basis for the rule-specificity account, assumes that a rule will always be used again on the next trial if it is successful (Ashby et al., 1998).

Interim Summary

In brief, young adults performed better than older adults at the categorization tasks and the two age groups had qualitatively different patterns of performance: For young adults, our data replicated the traditional pattern of accuracy (Type I > Type II > Type III = Type IV; e.g., Shepard et al., 1961). However, older adults showed superior performance in Type IV compared to Type II. These age differences were similar to those found by Rabi and Minda (2016) who hypothesized that older adults’ performance was driven by increased reliance on single-dimensional rules during learning. In the current study, statistical tests of single-dimensional rule use did not support this hypothesis.

Model-based Analysis

We presented the intuition above that constructing fewer clusters in the RMC (Anderson, 1991) would result in the observed age deficits. To verify that young and older adults did construct different numbers of clusters and that it could produce the same pattern of age deficits, we fit the RMC to the data. The RMC is a model that infers which items belong together in clusters, based on both their physical features and their category labels. In this model, the category label is treated as just another feature, so it is possible that items from two separate categories will be placed in the same cluster. When making category judgments, the RMC first finds the probability that the new item comes from each of the clusters (including the possibility that the item belongs in a new cluster) and then weights the prediction of each cluster/level of the category label by these probabilities.

The RMC used three parameters in its original formulation: a coupling parameter, *c*, a physical salience parameter, s_P, and a label salience parameter, s_L. The coupling parameter controls the prior probability of the number of clusters. A high coupling parameter means there will be fewer clusters, whereas a low coupling parameter means there will be more clusters. The two salience parameters control how "pure" each of the clusters are along the physical (e.g., size, form, and color) or label features, with lower values meaning that each cluster is more likely to contain only a single value of each feature (e.g., this cluster will only have triangles or only squares). For the label salience parameter, a low value means that it is less likely that two items from different categories will be placed in the same cluster. The RMC is also often augmented by a determinism parameter, *r*, which acts to bring response probabilities either closer to chance for low *r* or closer to deterministic performance for high *r* (Nosofsky, Gluck et al., 1994). Full details of the RMC are given in Appendix A.

To investigate which parameters were responsible for the differences between the age groups, we created a set of 16 models. Every model was fit using the same parameters for all participants within an age group, but the different models allowed for different sets of parameters to differ between groups. A description of each model along with several measures of how well each fit the data is shown in Table 1. For all of these measures, a lower value indicates a better model. The negative log likelihood was computed across all participants and only measures the fit to the data, while AIC and BIC adjust the overall negative log likelihood with penalties for model complexity. We also converted AIC and BIC values into the more interpretable AIC and BIC weights, which approximate the probability of each model given the data, assuming the models are equally likely before the experiment began (Akaike, 1978; Kass & Raftery, 1995; Wagenmakers & Farrell, 2004).

Using both AIC and BIC weights, the best model was clearly Model 14, which allowed for three of the four parameters to differ between young and older adults: s_P, c, and r. The performances of young and older adults predicted by this model are shown in Figure 6, and they generally match the human data well. The main discrepancy is that within each age group the model did not learn Type I tasks as quickly as participants did, but the overall accuracy predicted by the best-fitting parameters matched the ordering of accuracy on the problem types for each age group.

The best-fitting values of Model 14's parameters are shown in Table 2. Older adults had a higher best-fitting coupling parameter than did young adults, implying that they formed fewer clusters. However, unlike Davis et al. (2012), we allowed the physical and label salience parameters to vary, as these parameters can also affect the clustering of the stimuli. A more direct view of how young and older adults clustered the stimuli can be obtained by looking at assignments of items to clusters made by the model. We found that the different orders in which the trials were presented led to variability in the clusters formed across individuals with the same parameters. In Figure 7, we show the assignments made by the model for the last block of stimuli in the experiment. Whereas young and older adults both used two clusters for Type I, the model indicates that older adults were more likely to use fewer clusters to represent each of Types II-IV. Overall, older adults were not using as many clusters as were young adults.

The differences in parameters between young and older adults do not just impact how the items are clustered. These parameters also impact how a category judgment is made given a particular representation. Older adults had higher values of c, as well as lower values of s_P and r. For new items, the value of c controls the influence of the existing clusters relative to a new cluster that contains just the new item and thus has no label information. As a result, the higher value of c means that older adults have stronger category preferences than young adults given their representation. Relatedly, lower values of s_P for older adults mean that items will have a stronger match to clusters they belonged to in previous blocks, increasing the strength of category preferences. However, the lower value of r for older adults means that responses will be more stochastic and that the most likely category label will not be chosen as often.

To determine the overall impact of these parameter differences on how category labels are chosen for incoming items, we looked at what would happen if older adults clustered like older adults, but made choices like young adults. This was done to establish whether the predicted reversal in Type II and Type IV performance was due to the clustering of items or to the choice parameters. In essence we used different parameters at different stages of each trial: the young adults' parameters were used when making a category label prediction, but after receiving feedback the older adults' parameters were used to assign an item to a cluster. The impact of using the older adults' choice parameters on accuracy can be seen in Table 3. If older adults behaved like young adults while predicting category labels, then they perform equivalently or slightly better than young adults for Type I (because young and older adults used the same clusters), and perform better but not as well as young adults for Types II-IV. Importantly, the performance on Type IV problems is still predicted to be more accurate than on Type II problems, which means that the clustering, rather than the choice parameters, is controlling relative performance for these two problem types for older adults.

Beyond the accuracy on Types I-IV, we also investigated what the best-fitting version of the RMC predicted for the statistics we developed to test for the presence of single-dimensional rules. Predicted Hamming distances were calculated by finding the expected minimum distance to the set of single-dimensional rules for each block based on the model predictions for each stimulus. The predicted distances matched the empirical distances well, with the exception that participants corresponded to single-dimensional rules in the Type I task better than the model predicted. The Pearson correlation between model predictions and empirical distances across all participants, types, and blocks was 0.95.

Figure 5 shows the RMC predictions for consecutive maximally different stimuli. For the lower panel (“previous response correct”), the trials selected were just the same trials selected in the analysis of the data. These model predictions show the same overall patterns as the human data, in particular the near 0.5 rate of different responding for older adults in the Type II task. The Pearson correlation between the model and data across age groups and tasks for all responses was 0.96, and for previous response correct the correlation was 0.97.

Discussion

We investigated three hypotheses of the source of age differences in categorization in our experiment: rule-complexity, rule-specificity, and clustering. In line with Rabi and Minda (2016), our results supported an age-related reversal of performance: Type II task performance was reliably worse for older adults than Type IV task performance, but Type II performance was statistically the same (and numerically better) than Type IV for young adults.[3] Because the rule-complexity hypothesis predicts that Type IV performance would be impacted more than Type II performance for older adults, this effect serves as strong evidence against a rule-complexity explanation of age deficits. More generally, it is evidence against any explanation that holds that age deficits will always be larger when the task is more difficult. More subtly, we found that while Type III and IV performance was equivalent for young adults, older adults were perhaps slightly better at Type IV than III, providing some additional evidence against a rule-complexity account.

Rabi and Minda (2016) attributed the Type II deficit in older adults to rule-specificity. They argued that older adults were generally unable to learn complex verbal rules. They found very little evidence for perfect correspondence to single-dimensional rules in their data, and supposed that older adults were switching between single-dimensional rules as they received negative feedback on their performance. In our data, we also found evidence against older adults generally being able to learn complex verbal rules in our Hamming distance analysis. Furthermore, we looked closely at the data to see if single-dimensional rule use was plausible.[4] Our Hamming distance analysis provided additional evidence that single-dimensional rules were not being used consistently by older adults in the Type II task, as they did not appear to be moving closer to single-dimensional rules in that task. Also, our consecutive trial analysis showed that quickly switching between single-dimensional rules did not describe older adults well either. Older adults made responses consistent with using

[3] Type II has been found to be better than Type IV for young adults, but not in every experiment. The effect is more likely to be found under the instructions that we used (Kurtz et al., 2012).

[4] We did not fit a model of single-dimensional rule use because there are a variety of ways to implement this approach, so we made a more general qualitative argument.

the same single-dimensional rule only on about half of trials in which this behavior could be assessed, even when just looking at pairs of trials in which the first response was correct.

For a rule-based system, using a rule more often after receiving positive feedback on its performance is critical – otherwise no learning is occurring. Stronger assumptions have been made: the COVIS explicit rule-based system assumes that positive feedback always leads to using the same rule again on the next trial (Ashby et al., 1998). As a result, being inconsistent with a single-dimensional rule on half of trials after positive feedback is difficult to explain with a single-dimensional rule system, unless it is working extremely poorly: the system is randomly choosing amongst all possible rules on each trial with equal probability. However, we show in Appendix B that many participants responded reliably above chance. Together these results make for an argument against older adults using single-dimensional rules in the Type II task, where they showed the greatest deficits.

A remaining possibility for the rule-specificity hypothesis is that older adults were attempting to use multi-dimensional rules, but were just worse at finding the correct multi-dimensional rules compared to young adults. Our Hamming distance analysis argued against this interpretation because there was no trend toward older adults moving further away from single-dimensional rules over blocks, but there exist many different proposals of how complex rules are learned (e.g., COVIS, Rational Rules, or Nosofsky, Palmeri et al.'s, 1994, RULEX) and these would have to be examined in detail.

The clustering hypothesis better accounts for these age deficits. We formalized the clustering hypothesis in the RMC and quantitatively showed that the deficits can be explained as older adults being less able to form new clusters than young adults. The best-fitting RMC showed the expected reversal in Type II and Type IV performance between older and young adults. The RMC also matched the empirical data well on the Hamming distances and consecutive trial analysis that argued against single-dimensional rules.

Using the clustering hypothesis rather than rules to explain age deficits leads to reinterpretations of some past results. For example, Maddox, Pacheco, Reeves, Zhu, and Schnyer (2010) found equal age-related declines in rule-based and information-integration tasks (akin to Type II and Type IV, respectively) when both were generated from four clusters. If older adults struggle to produce as many clusters as young adults, these equal declines would be expected. Additionally, clustering can be used to explain some of the strongest evidence for rule-specificity age deficits: age-related increases in perseverative errors in the WCST (Rhodes, 2004). Clustering models can be used in associative learning tasks to explain how old associations do and do not interfere with new associations: if both old and new associations are part of the same cluster then there will be interference because they cannot be accessed separately, but if old and new associations are part of separate clusters then the new associations can potentially be accessed without interference (e.g., Gershman, Blei, & Niv, 2010). If older adults have more trouble creating new clusters, this then could explain why they show greater perseverative errors when the rule changes in the WCST.

Interpreting age declines as an increased difficulty in constructing new clusters yields a new interpretation of the relationship between working memory capacity and type of task. Rabi and Minda (2016) found that working memory capacity was related to performance on on Type II but not on Type I tasks. Instead of interpreting working memory capacity as related to performance on complex rules, we can interpret it as necessary for constructing more clusters, because each additional cluster means that there is more information to represent. Rabi and Minda argued that Type II may be more influenced by working memory than Type IV and recently, Stukken, Van Rensbergen, Vanpaemel, and Storms (2016) showed that higher working memory ability was related to utilization of a greater number of clusters during categorization. However, contrary to this view, Lewandowsky (2011) found that working memory capacity affected performance on Types I-VI similarly. More research is needed to clarify the relationship between working memory and clustering, especially as clustering is more naturally described as implicit memory, and older adults show greater deficits in explicit than implicit memory.

Clustering represents one implicit type of categorization, but there are others. Exemplar models, formalized as the Generalized Context Model (GCM; Nosofsky, 1984) and ALCOVE (Kruschke, 1992), also could potentially explain these results. These exemplar models use the mechanism of selective attention to produce better performance in Type II than Type IV problems. As it is easier to selectively attend to fewer dimensions, Type II has an advantage over Type IV because in Type II one of the dimensions can be completely ignored (see Figure 1). The claim for selective attention has been bolstered by findings that the performance advantage for Type II over Type IV only occurs for separable dimensions, like those used in our experiment, where selective attention can operate. The reverse pattern occurs with integral dimensions, such as the hue and saturation of colors, for which selective attention is much harder to employ (Nosofsky & Palmeri, 1996).

A deficit in selective attention is another non-rule-based approach for explaining the reversal of Type II and IV performance between young and older adults. However, it is not clear whether selective attention can explain our results. Maddox, Filoteo, and Huntington (1998) tested selective attention for integral and separable stimuli, investigating how well young and older adults could ignore irrelevant information on non-selected dimensions. They found that for separable dimensions, older adults were just as good as young adults at selective attention, though they only tested application of a known categorization rule rather than learning an unknown rule as we did in our experiment. Future work could combine our task with a selective attention task to see if individual differences in selective attention in older adults correspond with individual differences in Type II and Type IV performance.

In summary, we have demonstrated that utilization of fewer clusters in older adults provides a parsimonious account of age differences in the Shepard et al. (1961) categorization tasks. We argue that this view is more consistent with the data than a rule-complexity account, and a rule-specificity account that postulates a reliance on single-dimensional rules in older adults. This does not mean that older adults do not use single-dimensional rules: although the overall pattern of results was best explained by the RMC with age deficits in both cluster formation and choice, the RMC was not able to match the participant performance on Type I problems, which can be perfectly represented by single-dimensional rules. It could be that a hybrid model that combines single-dimensional rules and a clustering representation would better explain the data, or perhaps a hierarchical elaboration of the RMC that introduces rule-like behavior is needed (Heller, Sanborn, & Chater, 2009). The clustering hypothesis is a start, but there is much about categorization in older adults that still needs investigation.

References

Akaike, H. (1978). On the likelihood of a time series model. *The Statistician, 27,* 217-235.

Aldous, D. (1985). Exchangeability and related topics. In P. L. Hennequin (Ed.), *École d'été de probabilités de Saint-Flour, XIII—1983* (pp. 1-198). Berlin, Germany: Springer.

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

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

Bourne, L. E. (1970). Knowing and using concepts. *Psychological Review, 77*, 546-556.

Bruner, J. S., Goodnow, J. J., & Austin, G. A. (1986). *A study of thinking*. London: Transaction Publishers.

Cerella, J., Poon, L. W., & Williams, D. M. (1980). Age and the complexity hypothesis. In L. W. Poon (Ed.), *Aging in the 1980s: Psychological issues* (pp. 332-340). Washington, DC: American Psychological Association.

Chasseigne, G., & Lafon, P. (2002). Aging and rule learning: The case of the multiplicative law. *The American Journal of Psychology, 115,* 315-330.

Davis, T., Love, B. C., & Maddox, W. T. (2012). Age-related declines in the fidelity of newly acquired category representations. *Learning & Memory, 19*, 325-329.

Feldman, J. (2000). Minimization of Boolean complexity in human concept learning. *Nature, 407,* 630-633. doi: 10.1038/35036586

Filoteo, J. V., & Maddox, W. T. (2004). A quantitative model-based approach to examining aging effects on information-integration category learning. *Psychology and Aging, 19,* 171-182.

Filoteo, J. V., Maddox, W. T., Ing, A. D., Zizak, V., & Song, D. D. (2005). The impact of irrelevant dimensional variation on rule-based category learning in patients with Parkinson’s disease. *Journal of the International Neuropsychological Society, 11,* 503-513.

Gallistel, C. R. (2009). The importance of proving the null. *Psychological Review, 116*, 439-453. http://doi.org/10.1037/a0015251

Gershman, S. J., Blei, D. M., & Niv, Y. (2010). Context, learning, and extinction. *Psychological Review, 117*, 197-209.

Goodman, N. D., Tenenbaum, J. B., Feldman, J., & Griffiths, T. L. (2008). A rational analysis of rule-based concept learning. *Cognitive Science, 32*, 108-154. doi: 10.1080/03640210701802071

Griego, J. A., & Kliegel, M. (2007). Adult age differences in function concept learning. *Aging, Neuropsychology, and Cognition, 15*, 1-30.

Heller, K. A., Sanborn, A. N., & Chater, N. (December, 2009). *Hierarchical learning of dimensional biases in human categorization.* Paper presented at the Advances in Neural Information Processing Systems Annual Conference, Vancouver, Canada.

Hess, T. M. (1982). Visual abstraction processes in young and old adults. *Developmental Psychology, 18*, 473-484.

Hess, T. M., & Slaughter, S. J. (1986). Aging effects on prototype abstraction and concept identification. *Journal of Gerontology, 41,* 214-221.

Jarosz, A. F., & Wiley, J. (2014). What are the odds? A practical guide to computing and reporting Bayes factors. *Journal of Problem Solving, 7,* 2-9.

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association, 90*, 773-795.

Kurtz, J. K., Levering, K. R., Stanton, R. D., Romero, J., & Moris, 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, 32*, 552-572.

La Voie, D., & Light, L. L. (1994). Adult age differences in repetition priming: A meta-analysis. *Psychology and Aging, 9,* 539-553.

Light, L. L., Prull, M. W., La Voie, D. J., & Healy, M. R. (2000). Dual-process theories of memory in old age. In T. J. Perfect & E. A. Maylor (Eds.), *Models of cognitive aging* (pp. 238-300). New York: Oxford University Press.

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

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

Love, J., Selker, R., Marsman, M., Jamil, T., Verhagen, A. J., Ly, A., . . . Wagenmakers, E.-J. (2015). JASP (Version 0.6.5): https://jasp-stats.org/.

Maddox, W. T., Filoteo, J. V., & Huntington, J. R. (1998). Effects of stimulus integrality on visual attention in older and younger adults: A quantitative model-based analysis. *Psychology and Aging, 13,* 472-485.

Maddox, W. T., Pacheco, J., Reeves, M., Zhu, B., & Schnyer, D. M. (2010). Rule-based and information-integration category learning in normal aging. *Neuropsychologia, 48,* 2998-3008.

Maintenant, C., Blaye, A., & Paour, J.-L. (2011). Semantic categorical flexibility and aging: Effect of semantic relations on maintenance and switching. *Psychology and Aging, 26*, 461-466.

Mata, R., von Helverson, B., Karlsson, L., & Cüpper, L. (2012). Adult age differences in categorization and multiple-cue judgment. *Developmental Psychology, 48*, 1188-1201.

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

Minda, J. P., Desroches, A. S., & Church, B. A. (2008). Learning rule-described and non-rule-described categories: A comparison of children and adults. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 34*, 1518-1533.

Naveh-Benjamin, M., & Ohta, N. (2012). *Memory and aging*. New York: Taylor & Francis.

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

Nosofsky, R. M. (1991). Relation between the rational model and the context model of categorization. *Psychological Science, 2*, 416-421.

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

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

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

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

R Core Team (2016). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/

Rabi, R., & Minda, J. P. (2016). Category learning in older adulthood: A study of the Shepard, Hovland, and Jenkins (1961) tasks. *Psychology and Aging, 31*, 185-197.

Racine, C. A., Barch, D. M., Braver, T. S., & Noelle, D. C. (2006). The effect of age on rule-based category learning. *Aging, Neuropsychology, and Cognition, 13*, 411-434.

Raven, J. C., Raven, J., & Court, J. H. (1988). *The Mill Hill vocabulary scale*. London: H. K. Lewis.

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

Rhodes, M. G. (2004). Age-related differences in performance on the Wisconsin Card Sorting Test: A meta-analytic review. *Psychology and Aging, 19*, 482-494.

Rosch, E. H. (1973). On the internal structure of perceptual and semantic categories. In T. E. Moore (Ed.), *Cognitive development and the acquisition of language* (pp. 111-144). New York: Academic Press.

Rosseel, Y. (2002). Mixture models of categorization. *Journal of Mathematical Psychology, 46*, 178-210.

Rouder, J. N. (2014). Optional stopping: No problem for Bayesians. *Psychonomic Bulletin & Review*, *21*, 301-308.

Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., & Iverson, G. J. (2009). Bayesian t-tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin & Review*, *16*, 225–37. http://doi.org/10.3758/PBR.16.2.225

Salthouse, T. A. (2010). *Major issues in cognitive aging*. New York: Oxford University Press.

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

Sanborn, A. N., & Hills, T. T. (2014). The frequentist implications of optional stopping on Bayesian hypothesis tests. *Psychonomic Bulletin & Review, 21*, 283-300.

Shepard, R. N., Hovland, C. I., & Jenkins, H. M. (1961). Learning and memorization of classifications. *Psychological Monographs: General and Applied, 75*, 1-42.

Stukken, L., Van Rensbergen, B., Vanpaemel, W., & Storms, G. (2016). Understanding individual differences in representational abstraction: The role of working memory capacity. *Acta Psychologica, 170,* 94-102.

Vanpaemel, W., & Storms, G. (2008). In search of abstraction: The varying abstraction model of categorization. *Psychonomic Bulletin & Review, 15*, 732-749.

Vigo, R. (2006). A note on the complexity of Boolean concepts. *Journal of Mathematical Psychology, 50*, 501–510. http://doi.org/10.1016/j.jmp.2006.05.007

Vigo, R. (2009). Categorical invariance and structural complexity in human concept learning. *Journal of Mathematical Psychology, 53*, 203-221.

Wagenmakers, E-J. (2007). A practical solution to the pervasive problems of p-values. *Psychonomic Bulletin & Review, 14*, 779-804.

Wagenmakers, E-J., & Farrell, S. (2004). AIC model selection using Akaike weights. *Psychonomic Bulletin & Review, 11*, 192-196.

Wechsler, D. (1981). *Manual for the Wechsler adult intelligence scale — revised*. New York: Psychological Corporation.

West, R. L. (1996). An application of prefrontal cortex function theory to cognitive aging. *Psychological Bulletin, 120*, 272-292.

Wetzels, R., Matzke, D., Lee, M. D., Rouder, J. N., Iverson, G. J., & Wagenmakers, E.-J. (2011). Statistical evidence in experimental psychology: An empirical comparison using 855 *t* tests. *Perspectives on Psychological Science, 6*, 291-298.

Wittgenstein, L. (1953). *Philosophical investigations*. New York: MacMillan.

Table 1

Model Comparison Between Versions of the Rational Model of Categorization That Tests for Differences in Parameters Between Age Groups

Model	Parameters Differing Between Age Groups	Number of Parameters	Negative Log Likelihood	AIC	AIC Weights	BIC	BIC Weights
1	none	4	20740	41488	0.0000	41522	0.0000
2	s_P	5	20268	40547	0.0000	40589	0.0000
3	s_L	5	20517	41044	0.0000	41086	0.0000
4	c	5	20383	40776	0.0000	40819	0.0000
5	r	5	20339	40688	0.0000	40731	0.0000
6	s_P, s_L	6	20243	40498	0.0000	40549	0.0000
7	s_P, c	6	20305	40622	0.0000	40673	0.0000
8	s_P, r	6	20221	40453	0.0000	40504	0.0000
9	s_L, c	6	20324	40661	0.0000	40712	0.0000
10	s_L, r	6	20328	40668	0.0000	40719	0.0000
11	c, r	6	20238	40489	0.0000	40540	0.0000
12	s_P, s_L, c	7	20241	40495	0.0000	40555	0.0000
13	s_P, s_L, r	7	20205	40424	0.0000	40484	0.0000
14	**s_P, c, r**	**7**	**20178**	**40371**	**0.7136**	**40430**	**0.9943**
15	s_L, c, r	7	20200	40414	0.0000	40474	0.0000
16	s_P, s_L, c, r	8	20178	40373	0.2864	40441	0.0057

Note. s_P is the physical salience parameter, s_L is the label salience parameter, c is the coupling parameter, and r is the determinism parameter. Negative log likelihood is the goodness of fit of the model (smaller is better), and AIC and BIC are two different measures that balance goodness of fit with a penalty for model complexity (smaller is better). AIC Weights and BIC Weights transform the AIC and BIC values to approximate the probability of the model given the data (larger is better). The best model by both AIC and BIC is in bold.

Table 2

Best-fitting Rational Model of Categorization Parameters for Young and Older Age Groups

Age	s_P	s_L	c	r
Young	0.6888	0.1615	0.5044	0.7738
Older	0.4427	0.1615	0.7450	0.4540

Note. s_P is the physical salience parameter, s_L is the label salience parameter, c is the coupling parameter, and r is the determinism parameter. In the best-fitting model, s_L is the same for young and older age groups.

Table 3

Predicted Accuracy for Problem Types I-IV

Parameters	Type I	Type II	Type III	Type IV
Young adults	0.86	0.77	0.70	0.73
Older adults choosing like young adults	0.87	0.55	0.64	0.69
Older adults	0.80	0.54	0.60	0.63

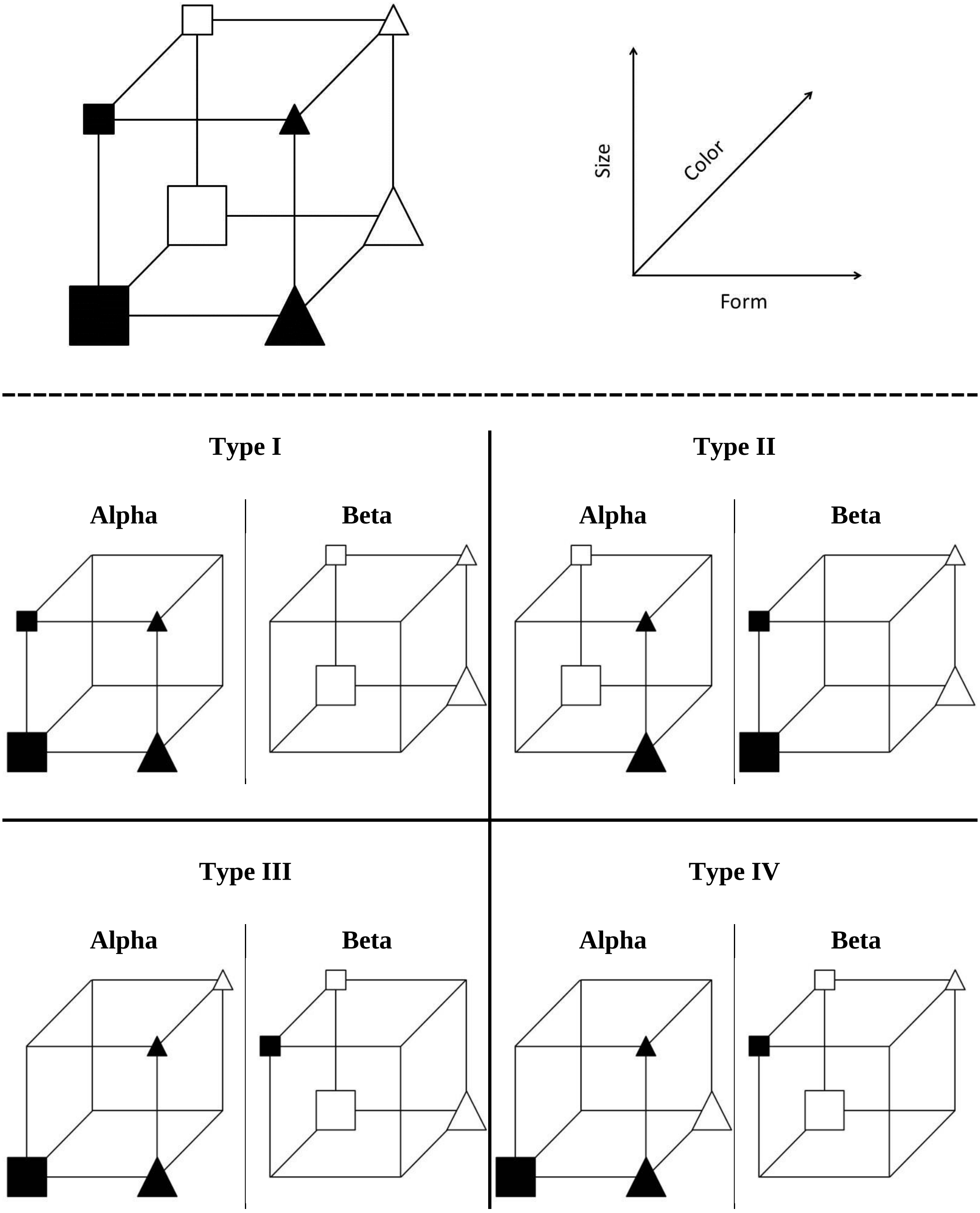

Figure 1. Top: Stimuli could vary along three dimensions (size, color, and form). Bottom: Examples of category membership for the eight shapes organized into two groups (alpha and beta) for the four categorization tasks (Types I to IV) used in the study.

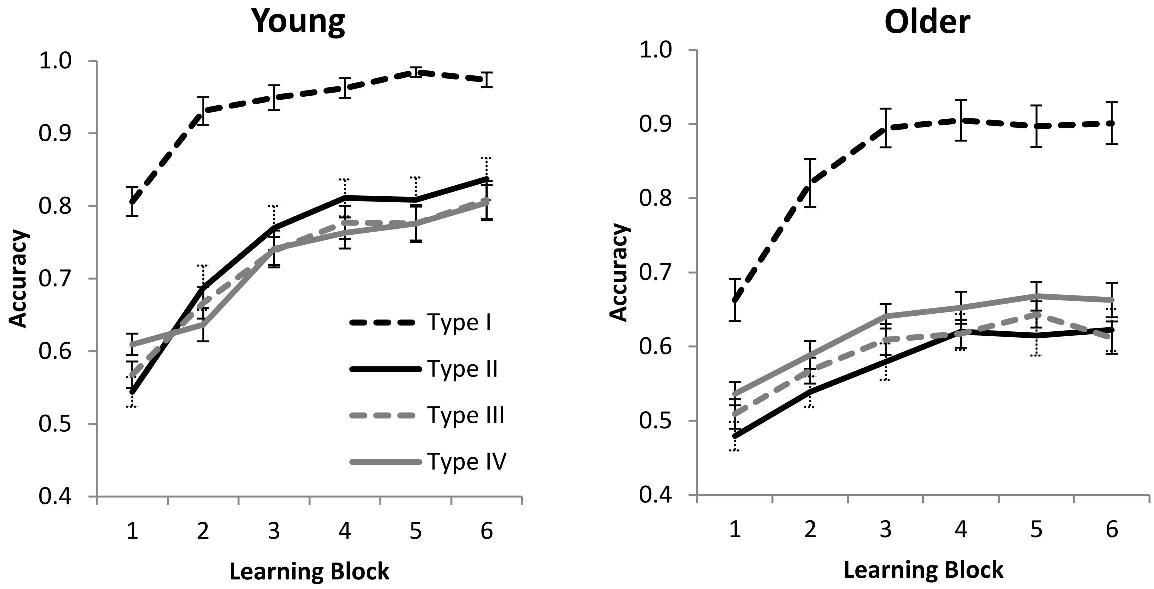

Figure 2. Accuracy for young and older adults learning categorization Types I, II, III and IV across six learning blocks (16 trials per block). Error bars are ±1*SE*.

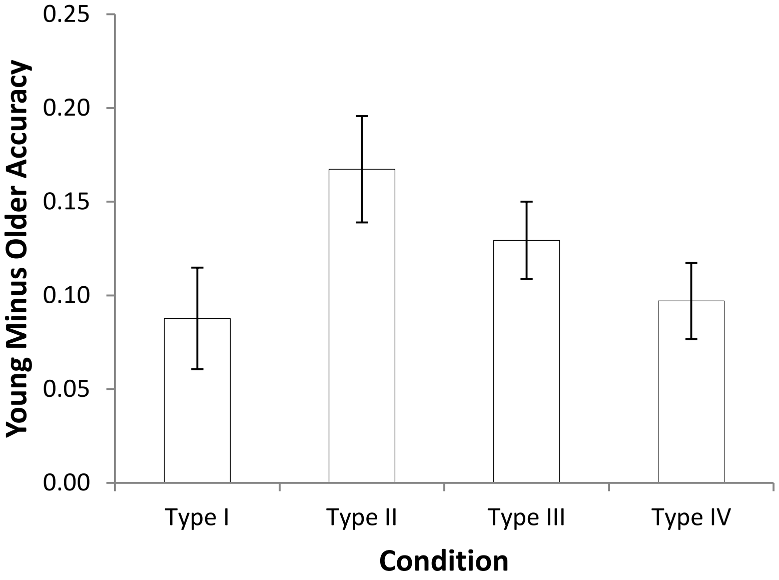

Figure 3. Magnitude of age deficits in learning for the different experimental conditions Types I to IV. Data are averaged across all six blocks. Error bars are ±1*SE*.

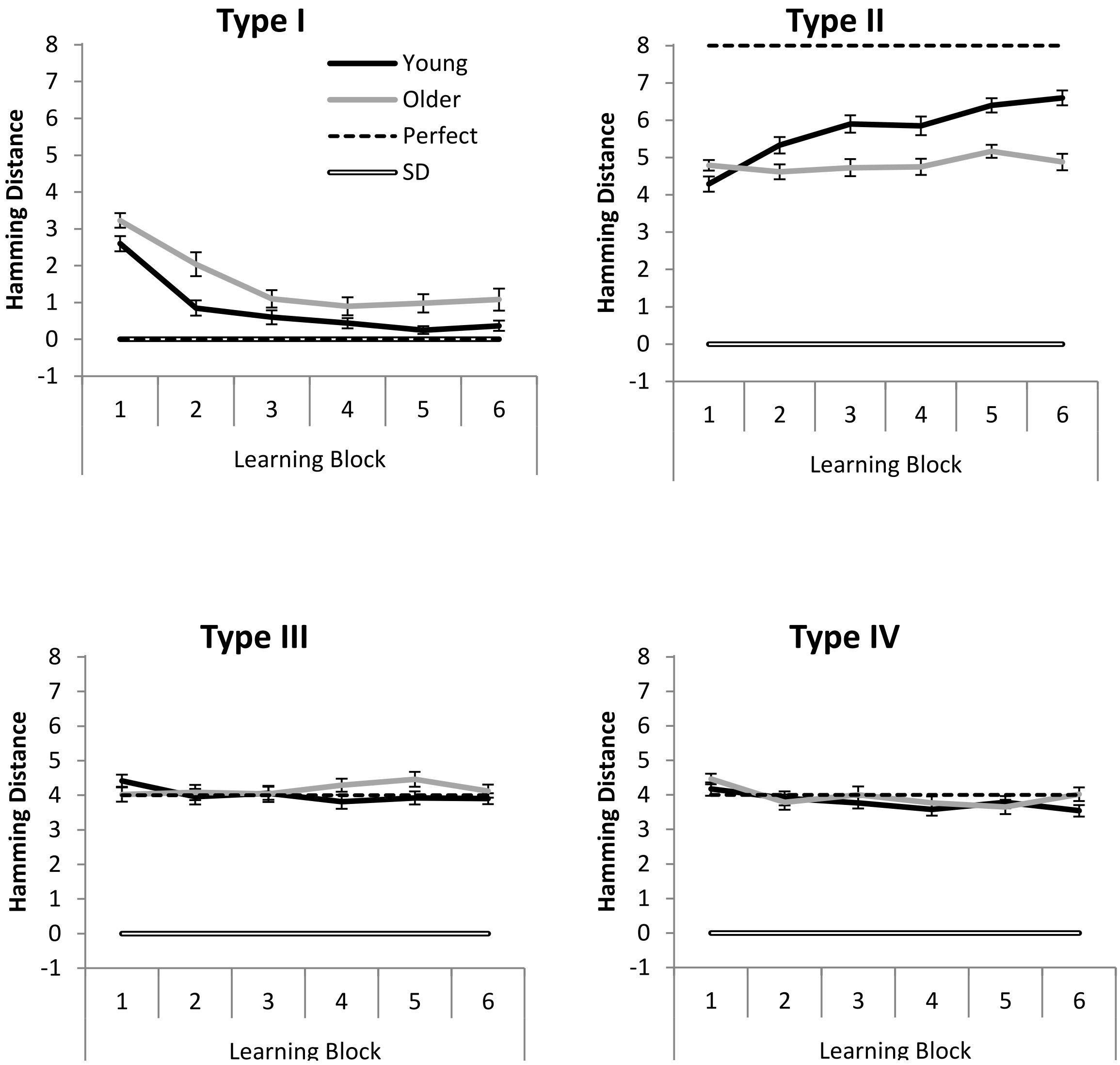

Figure 4. Hamming distance for young and older adults across learning blocks 1-6. The dashed line indicates the Hamming distance that would occur if participants were responding with 100% accuracy for each type, though this is necessary and not sufficient to produce perfect performance: matching this distance does not imply 100% accuracy. The hollow line indicates the distance corresponding to single-dimensional rule use (SD). Error bars are ±1*SE*.

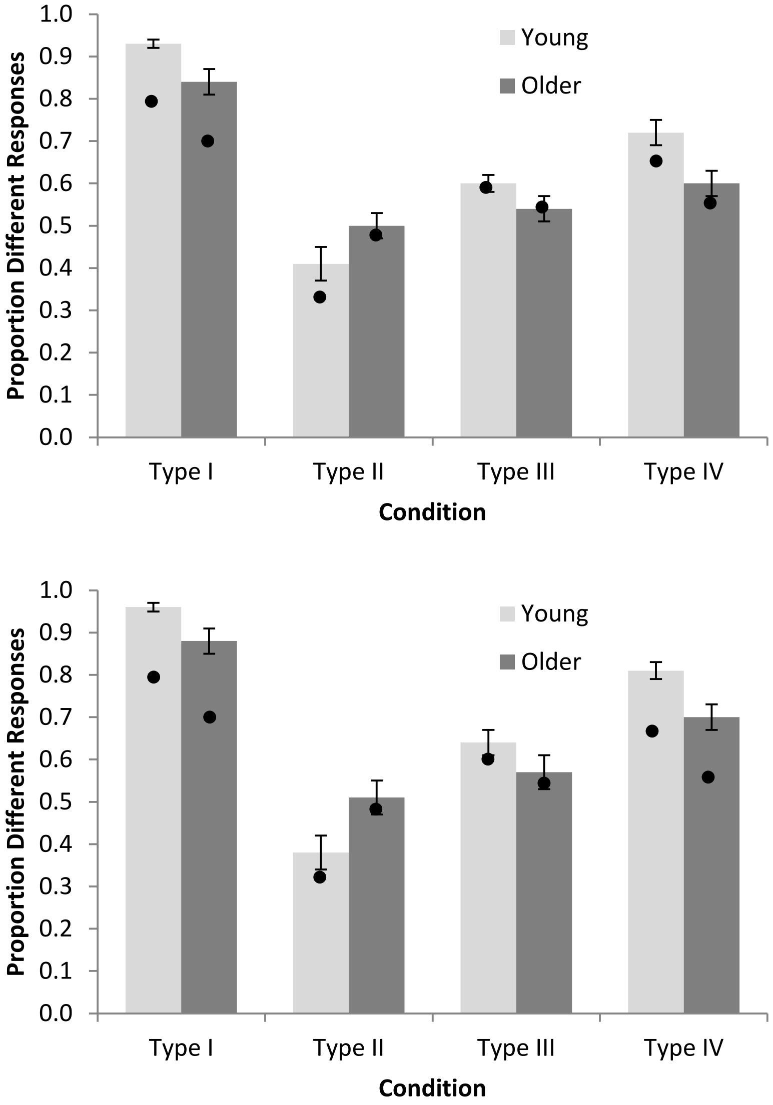

Figure 5. Proportion of different responses to all consecutive trials with maximally different stimuli (top panel), and to consecutive trials with maximally different stimuli where the previous response was correct (bottom panel). Error bars are ±1*SE*. Black circles indicate predictions from the Rational Model of Categorization.

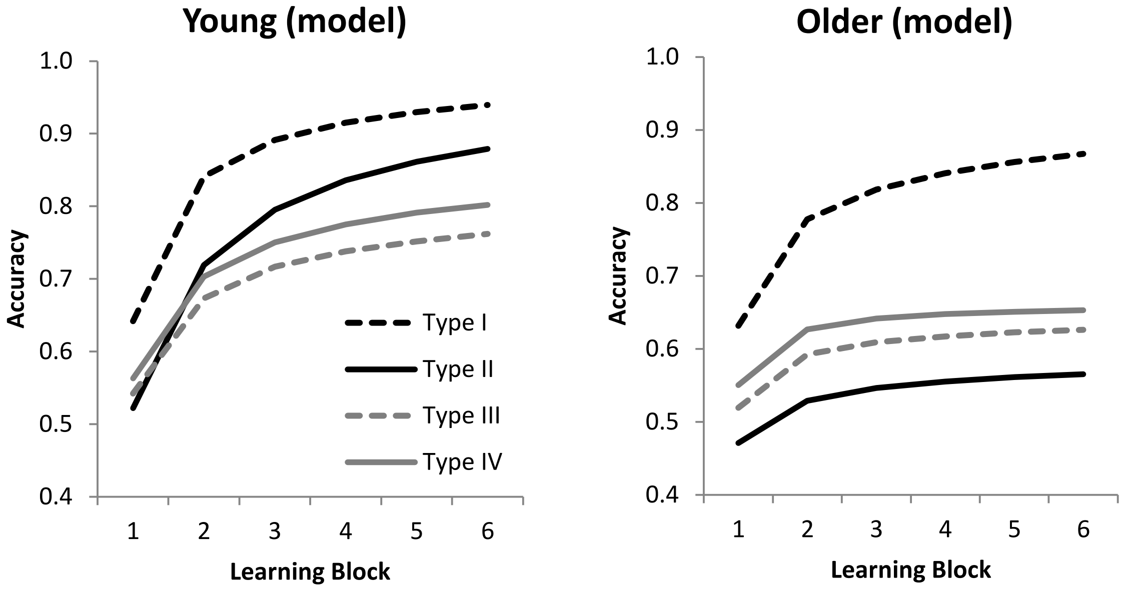

Figure 6. Model predictions for young and older adults learning categorization Types I, II, III and IV across six learning blocks.

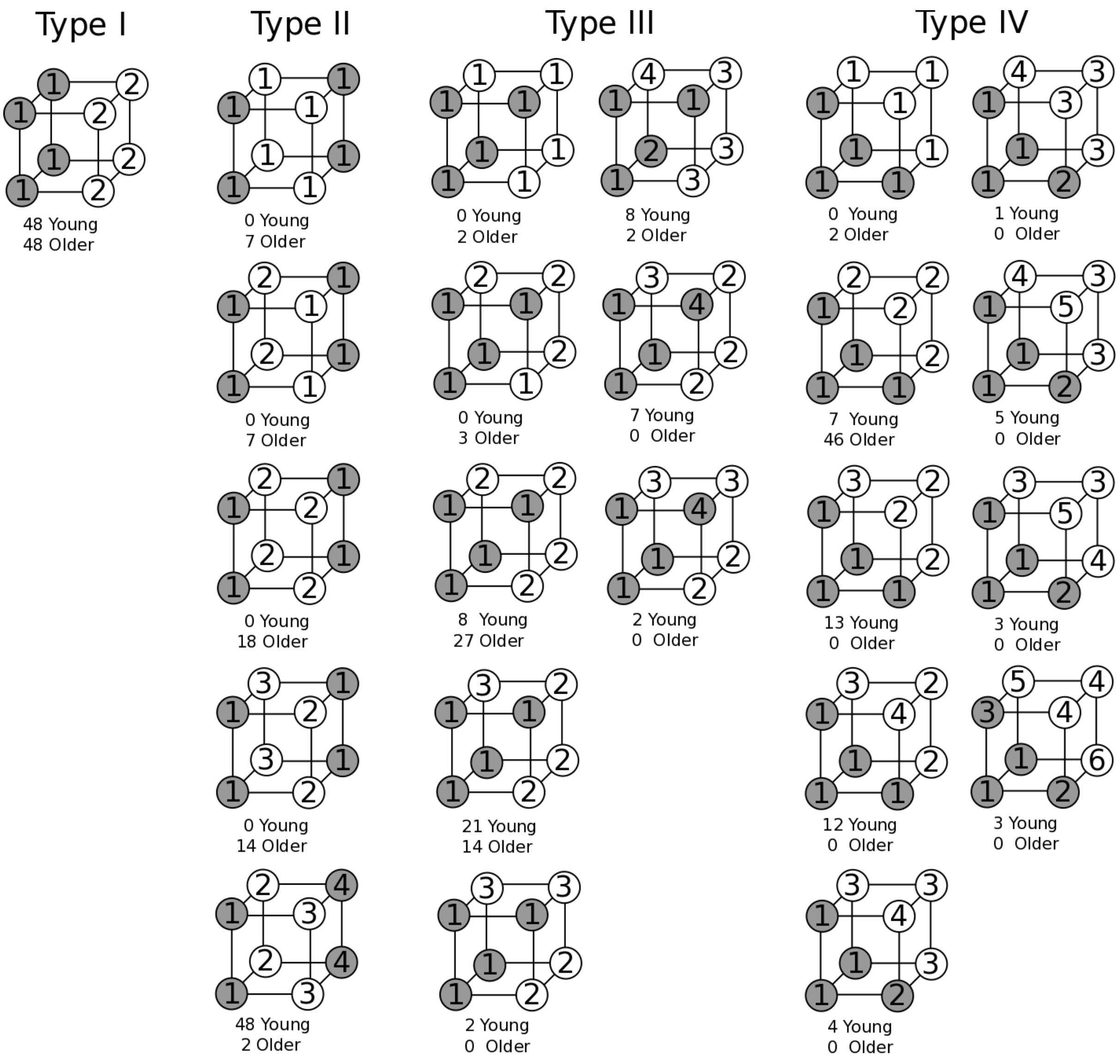

Figure 7. Visualization of how young and older adults clustered the items for Types I-IV. The plots underneath each type display the different ways the problem was clustered across participants. Within each plot, the three dimensions represent the three different physical dimensions, though the dimension identities have been ignored and cluster assignments renumbered to minimize the variety of different patterns. Gray and white circles indicate the feedback given to the items and the numbers within each circle label the cluster to which that item has been assigned. Text underneath each plot gives the number of young and older adults who used that set of clusters for that problem.

Appendix A: Details of the Rational Model of Categorization

When making a category judgment, the Rational Model of Categorization (RMC; Anderson, 1991) determines the probability that a new item, item n, belongs to the $i\mathrm{th}$ category ($y_n = i$), based on the physical features of that item, x_n, the physical features of all of the previously seen items, $\mathbf{x_{n-1}}$, and the category labels of all of the previously seen items, $\mathbf{y_{n-1}}$. This can be written as the probability of the category label and physical features of the new item, given all the previously seen items and labels

$$P(y_n = i|x_n, \mathbf{x_{n-1}}, \mathbf{y_{n-1}}) \propto P(y_n = i, x_n|\mathbf{x_{n-1}}, \mathbf{y_{n-1}}).$$

This probability is then transformed into the probability of making a binary category response i by raising it to the exponent r and then renormalizing

$$\frac{P(y_n = i|x_n, \mathbf{x_{n-1}}, \mathbf{c_{n-1}})^r}{P(y_n = i|x_n, \mathbf{x_{n-1}}, \mathbf{c_{n-1}})^r + (1 - P(y_n = i|x_n, \mathbf{x_{n-1}}, \mathbf{c_{n-1}}))^r}.$$

The probability of a new item and its category label is a weighted sum of the probability of the item and label arising from each of the clusters. Because the cluster indices are inferred from the data, these are marginalized out

$$P(y_n = i, x_n|\mathbf{x_{n-1}}, \mathbf{y_{n-1}}) = \sum_{\mathbf{z_n}} P(y_n = i, x_n|\mathbf{z_n}, \mathbf{x_{n-1}}, \mathbf{y_{n-1}})P(\mathbf{z_{n-1}}|\mathbf{x_{n-1}}, \mathbf{y_{n-1}})P(z_n|\mathbf{z_{n-1}}) \quad \text{(A1)}$$

where z_n is the cluster index of the cluster for item n, and $\mathbf{z_{n-1}}$ are the cluster indices for the previously seen items.

There are three terms on the right-hand side of Equation A1 that must be defined. First, the prior probability of the cluster index of the new item $P(z_n|\mathbf{z_{n-1}})$ is defined as a Chinese Restaurant Process prior which can flexibly interpolate between a single cluster for all the items and a different cluster for each item (Aldous, 1985). This prior can be written as a simple sequential rich-get-richer process

$$P(z_n = k|\mathbf{z_{n-1}}) = \begin{cases} \frac{cM_k}{(1-c)+c(i-1)} & \text{if } M_k > 0 \text{ (i.e., } k \text{ is old)} \\ \frac{1-c}{(1-c)+c(i-1)} & \text{if } M_k = 0 \text{ (i.e., } k \text{ is new)} \end{cases}$$

where M_k is the number of items assigned to cluster k and c is the coupling parameter which helps determine the number of clusters.

Second, the posterior on cluster assignments for previous items can also be built up as a sequential process

$$P(\mathbf{z_{n-1}}|\mathbf{x_{n-1}}, \mathbf{y_{n-1}}) \propto P(y_{n-1}, x_{n-1}|\mathbf{z_{n-1}}, \mathbf{x_{n-2}}, \mathbf{y_{n-2}})P(\mathbf{z_{n-2}}|\mathbf{x_{n-2}}, \mathbf{y_{n-2}})P(z_{n-1}|\mathbf{z_{n-2}}). \text{ (A2)}$$

We used the original inference algorithm for the RMC which assigns an item to the cluster that had the highest probability of producing that item. This approximation makes simulation feasible relative to summing Equation A1 over all possible partitions of the items into clusters, and it is deterministic which improves speed of computation. This approximation is sensitive to the ordering of the items, so the likelihood of each participant's responses was computed given the order in which he or she saw the items.

Finally, we need the likelihood of an item and its category label given the other items that are already assigned to a cluster. Because the RMC assumes that within a cluster the category label is independent from the physical features, and that all of the physical features are independent from one another

$$P(y_n = i, x_n|\mathbf{z_n}, \mathbf{x_{n-1}}, \mathbf{y_{n-1}}) = P(y_n = i|\mathbf{z_n}, \mathbf{y_{n-1}}) \prod_j P(x_{j,n}|\mathbf{z_n}, \mathbf{x_{n-1}})$$

where $x_{j,n}$ is the value of the $j\text{th}$ physical feature of item n.

The Shepard et al. (1961) learning task used three binary-valued physical features, so a Beta-binomial likelihood distribution was used

$$P(x_{j,n} = v|\mathbf{z_n}, \mathbf{x_{n-1}}) = \frac{B_v + s_P}{B_. + 2s_P}$$

where B_v is the number of items in the cluster that match the new item along the $j\text{th}$ physical feature, $B_.$ is the number of items in the cluster, and s_P is the parameter of a symmetric Beta prior distribution on the probability of obtaining different features. The likelihood of the binary-valued category feature uses the same form with a separate symmetric Beta prior

$$P(y_n = v|\mathbf{z_n}, \mathbf{y_{n-1}}) = \frac{B_v + s_L}{B_. + 2s_L} .$$

The parameters used to infer the category label and those used to infer the cluster assignments are defined by the model to be the same, but it is possible to use separate parameters in the two operations (e.g., Nosofsky, 1991). In the results, we separated the two processes by computing Equation A2 with one set of parameters, and then computing the response probabilities in Equation A1 (given the indices computed in Equation A2) using a different set of parameters.

Model Fitting Details

The likelihood of participant responses was determined using a Bernoulli likelihood. We implemented our model in R and used the nlm function from the stats package (R core team, 2016) to search for the best-fitting parameters. The parameter search was difficult because the predictions of the RMC can change suddenly with small changes in the parameters (e.g., Sanborn, Griffiths, & Navarro, 2010). We eventually settled on a procedure of running the nlm function until it converged and then restarting the algorithm by randomly jittering the best-fitting parameters. The random new starting point was a sample from a Gaussian distribution centered on the previous best fit with a standard deviation of 0.5% of the value of the parameter. We restarted the algorithm at least thirty times for each model, but finding that some of the models were still improving, we repeated this exercise until the log likelihood appeared stable and all models nested within more general models fit worse than the more general models. For the collection of models we analyzed, the parameter search required a month on a desktop computer.

Appendix B: Response Consistency

The experimental results showed that for older adults in the Type II task their responses matched between maximally different stimuli on approximately half of trials, even when they were correct on the previous trial. This pattern of behavior could be a result of random responding, or equivalently the result of choosing a new single-dimensional rule with equal probability on each trial from amongst the complete set of single-dimensional rules.

To determine if young or older adults were randomly responding in any of the tasks, we calculated a measure of response consistency, which can identify participants who are at chance accuracy but are still making consistent responses. This measure calculated for each stimulus the proportion of trials on which participants made the same response, which ranged from 0.5 (i.e., ‘Alpha’ to half the trials and ‘Beta’ to the other half) to 1 (i.e., responses were either always ‘Alpha’ or always ‘Beta’).

We simulated the distribution of response consistency for individuals performing according to chance. Ordering the simulated response consistency from lowest to highest, the 95th percentile was almost exactly 2/3, so this value was used as the cut-off for significance. The proportion of participants that were significantly above chance in response consistency is shown in Table B1 for each age group and task type.

Table B1

Proportion of Participants Significantly Above Chance in Response Consistency

	Type I	Type II	Type III	Type IV
Young	0.98	0.81	0.83	0.77
Older	0.90	0.46	0.54	0.79