When do learned transformations influence similarity and categorization?

Steven Langsford (steven.langsford@adelaide.edu.au)
School of Psychology, University of Adelaide

Andrew T. Hendrickson (d.hendrickson@tilburguniversity.edu)
Department of Communication and Information Systems, Tilburg University

Amy Perfors (amy.perfors@adelaide.edu.au)
School of Psychology, University of Adelaide

Daniel J. Navarro (d.navarro@unsw.edu.au)
School of Psychology, University of New South Wales

Abstract

The transformational theory of similarity suggests that when judging similarity, people are sensitive to the number of transformation operations needed to make two compared representations match. Although this theory has been influential, little is known about how transformations are learned and to what extent learned transformations affect similarity judgments. This paper presents two experiments addressing these questions, in which people learned categories defined by a transformation. In Experiment 1, when the transformations were directly visible, people had no trouble learning and applied their knowledge to similarity and categorization judgments involving previously unseen items. In Experiment 2, the task required transformations to be inferred rather than observed. People were still able to learn the categories, but in this more difficult case ratings were less strongly affected by training. Overall, this work suggests that newly learned transformations can impact similarity judgments but the salience of the transformation has a large impact on transfer.

Keywords: similarity; category learning; transformational similarity

Introduction

Calculating similarities is a core process in cognition (Medin, Goldstone, & Gentner, 1993) and plays a central role in categorization (Nosofsky, 1984). However, there is considerable debate about the fundamental building blocks for computing the similarity between objects that contain structured properties (Markman & Gentner, 1993; Hahn, Chater, & Richardson, 2003). One proposed basis for similarity is the transformational distance between items (Imai, 1977), which holds that the similarity between two objects is proportional to the number of steps required to transform one object into the other. Several papers outline the theoretical foundations of the approach (Chater & Vitányi, 2003; Chater & Hahn, 1997; Bennett, Gács, Li, Vitányi, & Zurek, 1998), the empirical evidence for it (Hahn et al., 2003; Hodgetts, Hahn, & Chater, 2009; Hahn, 2014), and the arguments against it (Larkey & Markman, 2005; Müller, van Rooij, & Wareham, 2009; Grimm, Rein, & Markman, 2012).

Transformation distances are sensitive to the primitive transformations available, but it is unclear how people might determine the relevant set (Grimm et al., 2012). Some transformations may be innate, but Müller et al. (2009) argue that for computational tractability, transformations must be organized in relatively small domain-specific sets. This suggests that where domain structure is learned, the relevant transformations for comparisons in that domain must also be learned.

We interpret the transformational approach as predicting a strong link between transformation learning and similarity judgments: learning a new transformation that directly connects two items should reduce the transformation distance between the items and thus increase the similarity between them. However, relatively little is known about how quickly transformations can be learned or how much new transformations impact similarity. The most relevant evidence comes from Hahn, Close, and Graf (2009), who found that people shown morphs from A to B rated similarity higher in the observed morph direction than the reverse direction. These results suggest that people are able to learn transformations over short timescales, and that there may be some impact on similarity. We extend this line of work using a transfer task, where test items are novel but instantiate the trained transformation. We manipulate whether transformations are directly observed or inferred, and separate measures of learning success from those of similarity judgment change.

Experiment 1

Can people learn categories that are defined by a novel transformation, and do they apply this transformation to novel categorization and similarity judgments? Experiment 1 addresses these questions with a training task designed to maximize the salience of a transformation relationship linking objects that belong to the same category. This is accomplished by showing the transformation after each categorization judgment during training. After training, we compare category membership and similarity judgments for a common set of previously unseen test items, contrasting responses from participants who were trained on different transformations.

Our results suggest that people learned the transformations and that this learning influenced subsequent categorization and similarity judgments. Items related by the newly-learned transformation were rated as more similar and more likely to belong to the same category. Items related by a novel transformation sharing some higher-level properties with the trained one were also rated as more similar and more likely to belong to the same category, although to a lesser extent.

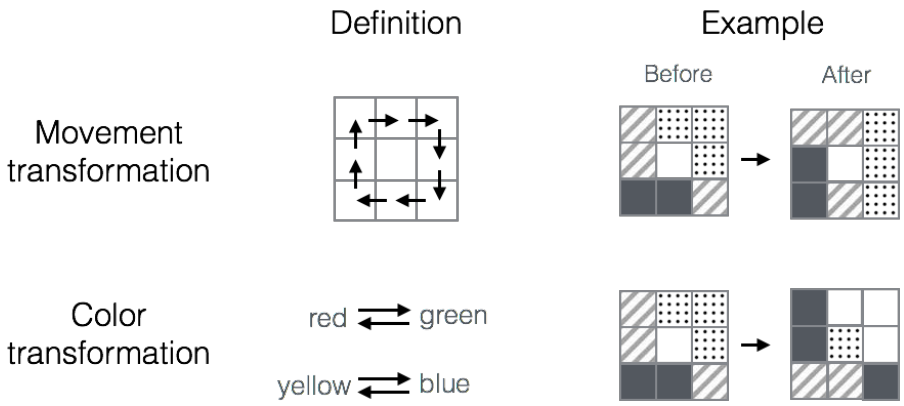

Figure 1: The two transformations used during the training phase of Experiment 1. In the MOVEMENT TRAINING people learned a non-rigid clockwise rotation transformation (top row), whereas in the COLOR TRAINING condition they learned a color swapping rule (bottom row). For both, the image on the left shows how that transformation was defined, and the image on the right gives an example on a particular stimulus. In this figure we use textures to display the four possibilities for each cell. The actual stimuli were presented in color, with the four possible values being red, green, yellow and blue.

Method

Participants Four hundred and forty-four participants were recruited via Amazon Mechanical Turk and paid US$0.75. 62% were male, with ages ranging from 18 to 67 (mean: 33.3). Three hundred and eleven participants were from the USA, 120 were from India, and 13 were from other countries. Forty-seven were excluded from all analyses: 12 for self-reported color-blindness and 35 for failing to pass check questions during the test phase of the experiment.

The experiment used two different pre-defined exclusion criteria, one based on training phase responses and one based on test phase responses. For the training phase, if any participant took more than 40 trials to learn any category that participant's data would be excluded. No participants were excluded on this basis. For the test phase, we also excluded any participant who gave an average similarity/categorization rating of less than 6 (out of 7) to the test trials with identical stimuli: 35 people were removed on this basis. One hundred and eighty six people were assigned to an IDENTITY condition in which the transformation to be learned was the identity transformation (i.e., no change). These participants easily learned the categories but were at floor for all generalization questions. Their results are not analyzed further.

Procedure The experiment consisted of six training phases and a test phase. Within each training phase, participants were trained on a new category of objects until their accuracy reached criterion. In the test phase, participants were asked to make categorization or similarity judgments of novel stimuli. All stimuli in the experiment consisted of 3x3 grids of colored cells, where each cell was a single color: red, yellow, blue or green (see the right panel of Figure 1). The stimuli were approximately 200 pixels wide on each side.

In each training phase, participants were shown a 'base' stimulus and told that it belonged to a category (e.g., *wugs*). Two items were displayed underneath with the question "Which of these is also a *wug*?" Participants were instructed to respond by clicking on the button located below their choice and were given feedback based on their choice. After an incorrect selection, the message "Sorry, try again" appeared and participants had to click the correct stimulus to proceed. After a correct selection, the message "Correct" appeared and an animation was presented morphing the base stimulus into the correct one. The next trial would then begin with the newly transformed item as the new target stimulus. For each category (e.g., *wugs*) this process continued until either the participant made four correct choices in a row or 40 trials had elapsed, at which point the experiment moved on to the next category (e.g., *philbixes*).

The set of stimuli in each category was determined by the base pattern and the transformation (shown in Figure 1). Each of the six training categories began with a unique 'base pattern' that was the same for all participants, and on each subsequent trial category members were generated by one application of the transformation. For participants in the COLOR TRAINING condition (n=114), the transformation from one item in the category to the next was a color-swapping rule in which cells that were colored red became green, green became red, blue became yellow, and yellow became blue. In the MOVEMENT TRAINING condition (n=144), the transformation that defined the set of items in the category was a non-rigid clockwise rotation of the cells in the grid. Applying this transformation caused the colors around the outside of the grid to shift one cell forward.

The test phase consisted of 20 test trials in which participants were asked to make judgments about pairs of novel stimuli that never appeared during training. The stimuli could be related to each other in one of six ways: identical (n=2), no simple relation (n=2), related by the trained movement (n=4) or trained color (n=4) transformations, or related by the novel movement (n=4) or novel color (n=4) transformations. The

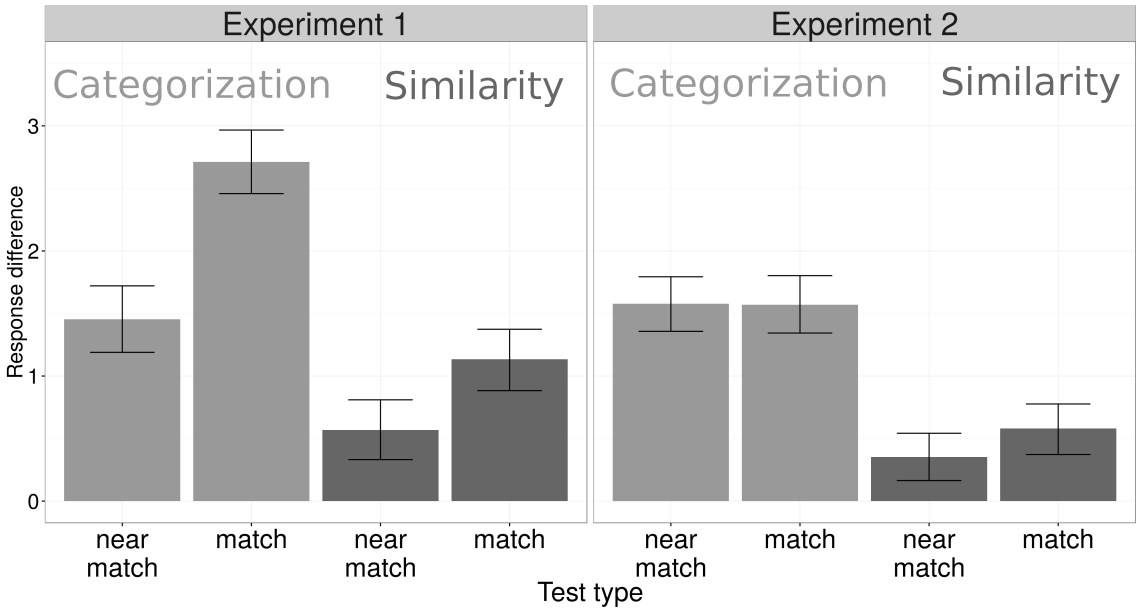

Figure 2: **Effect of transformation training**. The *y*-axis reflects the difference in responses given due to training condition, contrasting ratings given when test items do not match the training condition (NO MATCH) as compared to when they are related to the training, either as an exact MATCH or as a similar but novel NEAR MATCH. Thus, values above zero indicate effective training (in the case of MATCH) and generalization (in the case of NEAR MATCH). The left panel shows **Experiment 1**, which made the transformations explicit. In it, people learned and generalized the transformations for both categorization (light bars) and similarity (dark bars) questions, although the magnitude was smaller for similarity. The right panel shows **Experiment 2**, in which the transformations were less salient. In that case, learning and generalization were evident for categorization questions, but these were much larger than for similarity. Error bars express 95% credible intervals for a Bayesian t-test.

basis for these relations were not equally available to all participants: test items instantiating color transformations were unrelated for people given the movement training, and vice versa, manipulating the relation of the test items to the training while keeping the items themselves constant. The identical and no simple relation trials were of the same form as the test trials but only used for attention-check exclusions and not analyzed further. The novel movement transformation consisted of shifting all cells in the grid down by one row and moving the bottom row to the top. The novel color transformation swapped red with blue and green with yellow.

The order of test trials was randomized. Half the participants in each condition were asked to make CATEGORIZATION judgments by rating how likely it is that the two stimuli "have the same name" from "Not at all" to "Extremely" on a seven point scale. The other half were asked to rate the SIMILARITY of the two stimuli on a seven point scale.

In summary, there were two training conditions, each using a different transformation. There were four critical types of test item (excluding attention checks). The critical property of interest was the relationship between the test item and the training condition: did the test reflect the same or similar transformation as the training? The same items had different status for different participants depending on the training they saw: test items were considered to be MATCH trials when the two stimuli being compared were related by an application of the trained transformation, NEAR MATCH trials when the test stimuli were related by a transformation similar but not identical to the trained one, and NO MATCH when the test stimuli were not related to the training. Thus, for a person who received COLOR TRAINING, a test item involving that same color transformation would be a MATCH, one involving the novel color transformation would be a NEAR MATCH, and the two movement-related test items would be NO MATCH. None of the test items were previously seen in training.

Results

We first wish to establish whether the training phases were of comparable difficulty. We therefore looked at both how fast people reached the mastery criterion as well as the exclusion rates between conditions. People reached the criterion of four correct responses in a row in an average of 6.3 trials in the MOVEMENT TRAINING condition, and 5.8 in COLOR TRAINING, with 95% of all categories learned in eight trials or less. Participant inclusion rates were also comparable across conditions, at 86.0% and 87.5%, as was average accuracy over all trials (85% and 88% in the MOVEMENT and COLOR training respectively). This suggests that people learned to effectively distinguish category members from foils and that both transformations were similarly difficult.

To test the impact of training on people's ratings, we examined the degree to which their responses were different on the MATCH and NEAR MATCH test items from the NO MATCH baseline. NO MATCH baseline ratings for test items related

SIMILARITY Judgments				
Test item relationship	Test item average	NO MATCH item average	Difference	BF
MATCH items	3.61 (1.71)	2.49 (2.00)	1.13	> 1000
NEAR MATCH items	3.14 (1.71)	2.58 (1.91)	0.569	> 1000

CATEGORIZATION Judgments				
Test item relationship	Test item average	NO MATCH item average	Difference	BF
MATCH items	4.17 (1.76)	1.46 (1.83)	2.71	> 1000
NEAR MATCH items	3.15 (2.11)	1.7 (1.922)	1.45	> 1000

Table 1: **Descriptive statistics and hypothesis tests for Experiment 1**. For each of the MATCH and NEAR MATCH items (first column), we show the average responses for each (second column) compared to the NO MATCH baseline on the same items (third column). We performed a Bayesian t-test on the difference between these (fourth column) and found that in all cases there was a strong effect of training (fifth column).

by a COLOR transformation came from participants exposed to MOVEMENT TRAINING, baseline ratings for test items related by a MOVEMENT transformation came from participants exposed to COLOR TRAINING. In both cases the stimuli involved in the MATCH and contrasting NO MATCH groups were physically identical, likewise for NEAR MATCH items and their corresponding NO MATCH group. The left panel of Figure 2 illustrates these differences due to training experience. For instance, the MATCH bar reflects the difference between responses for the same item in the MATCH and NO MATCH conditions (thus, a value higher than zero indicates that the transformation training had an effect). Similarly, the NEAR MATCH bar reflects the difference between responses for the same item in the NEAR MATCH and NO MATCH conditions (thus, a value higher than zero indicates some generalization of training to a similar transformation).

Table 1 shows the absolute responses for the items of interest (i.e., the MATCH or NEAR MATCH items, in the second column) and the unrelated NO MATCH items in the third column. We used a Bayesian t-test (Morey, Rouder, & Jamil, 2014; Rouder, Speckman, Sun, Morey, & Iverson, 2009) to quantify the difference between them (fourth column), yielding a Bayes factor associated with the size of that different (fifth column). There was a strong ($BF > 1000 : 1$) effect of training for both the categorization and similarity judgments. However, these two types of judgment were impacted to different extents. For instance, the overall difference in item ratings between training conditions was between 1.07 and 1.33 larger (95% credible interval) for categorization judgments than similarity judgments.

Similarly, both MATCH and NEAR MATCH test item ratings differed strongly due to training ($BF > 10^3 : 1$), but to different extents. For instance, the difference due to training was between 0.67 and 0.93 rating points larger for MATCH as opposed to NEAR MATCH transformations. This suggests that people were less likely to generalize their responses as strongly to similar but not identical transformations.

Conclusion

The results of Experiment 1 show that learning categories that are defined by a transformation can lead people to produce consistently different patterns of judgments for novel items. Test items that were connected either by a learned transformation or a similar transformation were reliably rated higher. This increase in rating was found for similarity judgments as well as judgments about category membership.

This pattern of results is consistent with the predictions of the transformational account of perceptual similarity (Hahn et al., 2003). Furthermore, it suggests that by learning categories that are related by a transformation people can infer the transformation and apply it to novel items and categories.

That said, it is unclear to what extent the training in Experiment 1 is reflective of real-world transformation learning. In the experiment, objects were shown transforming into each other repeatedly; but in the real world, many transformations that define categories occur at a time scale that people cannot directly observe (e.g. seasons, aging, etc.). Experiment 2 aimed to test if the explicit presentation of the transformation was necessary to elicit quick learning and generalization of transformations.

Experiment 2

Experiment 1 provides "in principle" evidence that people are capable of learning rich knowledge about classes of stimulus transformations and the categories to which they are applicable. However, the structure of our task made learning as easy as possible: during the training phase participants were explicitly shown the transformation at the end of every trial. When learning new categories in real life it is more typical for people to encounter a variety of exemplars. For example, when learning the transformations involved in the aging of human faces, people observe many faces at different ages, but do not directly observe the aging process. It is thus unclear how generalizable these results are.

This issue is particularly important for evaluating the transformational account of similarity. With a few notable exceptions, such as rotation, the majority of transformations plausibly involved in comparisons are unobservable. Experiment 2 addresses the question of how easily learnable transformations when they are more implicit. By increasing the difficulty of the task, this manipulation also allowed us to examine the extent to which variation in ease of transformation learning is reflected in similarity.

Method

Participants Two hundred and fifty-two participants were recruited via Amazon Mechanical Turk and paid US$1. 60% were male, with ages ranging from 19 to 67 (mean: 34.7). Two hundred and forty-seven participants were from the USA, with the remainder from India, South America, and the UK. Fifty-three were excluded from all analyses: 2 for self-reported color-blindness, 12 for not completing the experiment, and 39 for failing exclusion criteria.

The experiment used two different predefined exclusion criteria, one based on training phase responses and one based on test phase responses. For the training phase, if any participant took more than 30 trials to learn two of the last three categories that participant's data would be excluded (this number was arrived at based on pilot data). Twelve participants were excluded on this basis. For the test phase, any participant who gave an average similarity/categorization rating of less than 6 (out of 7) to the identical trials were excluded: Twenty-seven people were removed on this basis. Ninety-five participants were in a COLOR TRAINING condition and 92 were in a MOVEMENT TRAINING condition. Sixty-five participants were in an IDENTITY condition and their results are not analyzed further.

Procedure

As in Experiment 1, this experiment consisted of six training phases and a test phase. Within each training phase, a new category of objects was learned until a mastery criterion was reached. In the test phase, participants were asked to make categorization and similarity judgments of novel stimuli. However, a number of aspects of the experiment differed from Experiment 1.

Based on pilot testing of category learning, the stimuli were simplified by adding the constraint that each stimulus contained at least six cells that shared the same color. Furthermore, the COLOR TRAINING transformation was modified to increase the number of possible stimuli within the categories. Instead of changing all colors (red to green, green to red, yellow to blue, and blue to yellow) as a single transformation, this was broken into two transformations. A single transformation consisted of either swapping the colors of red and green, or swapping yellow and blue. This doubled the number of stimuli in each condition in the COLOR TRAINING condition to more closely match the number in the MOVEMENT TRAINING condition.

The structure of the training trials also differed from Experiment 1. On each trial participants were shown two stimuli and asked if both items belonged in the category. There was always at least one category member displayed. In half the trials, the other stimulus was also in the category and related by one application of the transformation being trained. In the other half, the other stimulus was not in the category. After participants responded yes or no they were given feedback indicating if they were correct, but unlike in Experiment 1 they did not observe the actual transformation.

Participants proceeded to the next category when they were correct on 8 of 10 trials. Consecutive sets of six trials were constrained to contain three 'yes' and three 'no' trials (in shuffled order), meaning participants reaching criterion answered both 'yes' and 'no' correctly. The test phase was largely similar to Experiment 1 except that the UNRELATED trials were removed and trials were grouped into four blocks, with order of presentation randomized within each block to avoid runs of similar test items.

Results

The results indicate that difficulty was higher than Experiment 1, but comparable across conditions. People reached the accuracy criterion in an average of 14.12 trials in COLOR TRAINING and 13.8 trials in MOVEMENT TRAINING. Inclusion rates were comparable between conditions at 74% and 80% respectively.

As in Experiment 1, we were interested in whether responses to test items were different based on whether the transformation involved was a MATCH, NEAR MATCH, or NO MATCH to the trained transformation. The right panel of Figure 2 shows the differences in responses, analogous to the same analysis in Experiment 1, with the associated Bayesian t-test results shown in Table 2. In all cases we find strong evidence that participants' ratings for the same items were higher when they had a MATCH or NEAR MATCH relationship to training as opposed to NO MATCH status ($BF > 49:1$ at minimum). This suggests that training was effective and people were capable of learning the transformations even if they were not explicitly shown.

That said, the size of the difference depended on question type. Category learning showed a much larger effect: the difference was between 1.09 and 1.31 points larger (95% credible interval) for categorization questions than similarity ones. Unlike in Experiment 1, MATCH and NEAR MATCH status were not strongly differentiated: the difference involving MATCH status items as opposed to NEAR MATCH status items plausibly included zero (with a 95% CI between -0.01 and 0.21).

General Discussion

The results from Experiment 1 showed that people are capable of learning a novel transformation, recognizing that this transformation is relevant to determining category membership, and applying the learned transformation when assessing similarity between items belonging to novel categories. This finding is consistent with the learning effect seen in Hahn et al. (2009), but extends previous results in showing systematic generalization across related transformations.

Experiment 2 echoes these results and further finds that the effect is not limited to training in which people see objects transforming; seeing labeled category members can induce a change in judgments. However, there are two notable differences from Experiment 1. First, as the transformations become less prominent during training, they seem to have less impact on subsequent judgments, particularly for similarity.

		SIMILARITY Judgments		
Test item relationship	Test item average	NO MATCH item average	Difference	BF
MATCH items	3.56 (1.6)	2.98 (1.75)	0.58	> 1000
NEAR MATCH items	2.66 (1.53)	2.31 (1.62)	0.35	49

		CATEGORIZATION Judgments		
Test item relationship	Test item average	NO MATCH item average	Difference	BF
MATCH items	3.72 (1.85)	2.14 (2.15)	1.57	> 1000
NEAR MATCH items	2.93 (1.98)	1.35 (1.82)	1.58	> 1000

Table 2: **Descriptive statistics and hypothesis tests for Experiment 2**. For each of the MATCH and NEAR MATCH items (first column), we show the average responses for each (second column) compared to the NO MATCH baseline on the same items (third column). We performed a Bayesian t-test on the difference between these (fourth column) and found that in all cases there was a strong effect of training (fifth column).

Second, the novel and trained transformations were less well differentiated.

The attenuation of the training effects seems likely to be a result of task difficulty, with people less inclined to shift their judgments based on training that was less clear. However the lack of differentiation between trained and novel but similar transformations is harder to interpret. Possibly participants formed an incomplete representation of the transformation which was applicable to both near and exact matches to training, but the form of this representation is unclear. People's success in distinguishing targets from foils at training suggests they did not simply track which features remain invariant (e.g., noting in the MOVEMENT TRAINING condition that colors are preserved and in the COLOR TRAINING condition that configurations are preserved).

In terms of the predictions of transformational similarity, our results are somewhat mixed. It is clear that people learn transformations relevant to a new domain quickly, and that such transformations can be applied to categorization and similarity judgment. However, the pattern of generalization between exact matches and near-matches would seem to require some kind of graded availability of transformations based on family resemblances between them, complicating the computation of transformation distances.

Transformations as features are common in natural categories, for example growth and aging or characteristic movement. Despite this, their role in similarity judgments over structured representations remains unclear. Taking as a starting point predictions implied by tractability constraints on the transformational account of similarity, the two studies presented here examine the conditions under which transformation learning might influence similarity and categorization. Our results show that people can learn transformations quickly and use them in subsequent similarity and categorization judgments. However, productive use of the transformations depends to some extent on the ease with which the transformation was learned, and in both easy and difficult learning conditions involves generalization across related transformations.

Acknowledgments

SL was supported by an Australian Government Research Training Program Scholarship. DN received salary support from ARC grant FT110100431 and AP from ARC grants DP110104949 and DP150103280.

References

Bennett, C. H., Gács, P., Li, M., Vitányi, P. M., & Zurek, W. H. (1998). Information distance. *Information Theory, IEEE Transactions on*, *44*(4), 1407–1423.

Chater, N., & Hahn, U. (1997). Representational distortion, similarity and the universal law of generalization. In *Simcat97: Proceedings of the interdisciplinary workshop on similarity and categorization.*

Chater, N., & Vitányi, P. (2003). The generalized universal law of generalization. *Journal of Mathematical Psychology*, *47*(3), 346–369.

Grimm, L. R., Rein, J. R., & Markman, A. B. (2012). Determining transformation distance in similarity: Considerations for assessing representational changes a priori. *Thinking & Reasoning*, *18*(1), 59–80.

Hahn, U. (2014). Similarity. *Wiley Interdisciplinary Reviews: Cognitive Science*, *5*(3), 271–280.

Hahn, U., Chater, N., & Richardson, L. B. (2003). Similarity as transformation. *Cognition*, *87*(1), 1–32.

Hahn, U., Close, J., & Graf, M. (2009). Transformation direction influences shape-similarity judgments. *Psychological Science*, *20*(4), 447–454.

Hodgetts, C. J., Hahn, U., & Chater, N. (2009). Transformation and alignment in similarity. *Cognition*, *113*(1), 62–79.

Imai, S. (1977). Pattern similarity and cognitive transformations. *Acta Psychologica*, *41*(6), 433–447.

Larkey, L. B., & Markman, A. B. (2005). Processes of similarity judgment. *Cognitive Science*, *29*(6), 1061–1076.

Markman, A. B., & Gentner, D. (1993). Structural alignment during similarity comparisons. *Cognitive psychology*, *25*(4), 431–467.

Medin, D. L., Goldstone, R. L., & Gentner, D. (1993). Respects for similarity. *Psychological review*, *100*(2), 254.

Morey, R. D., Rouder, J. N., & Jamil, T. (2014). Bayesfactor: Computation of bayes factors for common designs [Computer software manual]. Retrieved from `http://CRAN.R-project.org/package=BayesFactor` (R package version 0.9.8)

Müller, M., van Rooij, I., & Wareham, T. (2009). Similarity as tractable transformation. In *Proceedings of the 31st annual conference of the cognitive science society* (pp. 50–55).

Nosofsky, R. M. (1984). Choice, similarity, and the context theory of classification. *Journal of Experimental Psychology: Learning, memory, and cognition*, *10*(1), 104.

Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., & Iverson, G. (2009). Bayesian t tests for accepting and rejecting the null hypothesis. *Psychonomic bulletin & review*, *16*(2), 225–237.