Violations of the Local Independence Assumption in Categorization

Jesse Zee (jesse.zee@student.kuleuven.be)
Gert Storms (gert.storms@ppw.kuleuven.be)
Steven Verheyen (steven.verheyen@ppw.kuleuven.be)
Faculty of Psychology and Educational Sciences, University of Leuven
Tiensestraat 102, B-3000 Leuven, Belgium

Abstract

The Threshold Theory of categorization posits that prior to making a categorization decision a respondent assesses the similarity between item and category and compares it against a personal threshold. Only if the item-category similarity exceeds this threshold should the respondent endorse the item as a category member. The Threshold Theory thus assumes that a single latent variable, item-category similarity, suffices to explain all categorization patterns. We put this assumption, known as local independence, to test by providing respondents with sets of four items to categorize. These items are equated in terms of their similarity to the target category, but are by design comprised of two pairs of similar items, such as *rollerskates-skateboard* and *horse-mule* for the category of VEHICLES. Contrary to the local independence assumption, the items within the pairs are more likely to receive the same categorization decision than the items across pairs. We explain how these categorization patterns can be accommodated within the Threshold Theory framework, either by giving up the assumption that everyone assesses item-category similarity in the same way for differently weighted category information, or by explicitly incorporating item-item similarity in addition to item-category similarity.

Keywords: categories; concepts; similarity; typicality; Threshold Theory; vagueness.

Introduction

Perceptions, thoughts, beliefs and other such mental phenomena are the product of the constant interaction of the embodied brain with the world. Indeed, the intentionality and meaning of these mental phenomena are most commonly about the outside world, as is the case with perception (Koendrink, 1999). Analogously, mental representations of categories and items are assumed to exist in our minds, to such an extent that we cannot trace the private meaning a person entails of any word. We can merely derive the mathematical models that best describe the patterns and behavior of such mental phenomena. Furthermore, the private nature of language and differences in brain and environment bring about the individually differential responses observed in many psychological studies and particularly in studies of vague concepts: concepts that specify categories without clearly defined boundaries.

To elucidate the categorization decisions of people when confronted with categorization tasks involving such vague concepts, most accounts ground their theory on the notion of similarity. Categorization is assumed to occur when item and category are sufficiently similar. Hampton (1998), for instance, investigated the categorization of instances at the border of vague categories. He found that for the majority of borderline cases, the probability of being placed in the category was directly predictable from how typical or representative the items were of the category (a proxy of item-category similarity).

The Threshold Theory of categorization (Hampton, 2007) offers a process account of how item-category similarity is used to arrive at categorization decisions. It posits that individuals verify whether the similarity meets a personal threshold in order to include the item in the category. Like most accounts of categorization, the Threshold Theory assumes local independence: the decisions pertaining to two items are considered conditionally independent of each other given their item-category similarity values. Any relationship that might exist between the categorization decisions should be explained by these similarity values. In other words, the theory invokes a latent variable, item-category similarity, that is believed to be sufficient to explain the entire range of categorization patterns one might observe.

Verheyen and Storms (2011) used a formal procedure to demonstrate violations of the local independence assumption in semantic categorization. They applied a formalization of the Threshold Theory to categorization data from eight natural language categories and found unexplained structure in the residuals. This result shows that item-category similarity does not suffice to account for all categorization patterns. Verheyen and Storms continued to show that many of the dependencies between items that they observed could be accounted for by item-item similarity. In categorizing items as FURNITURE or not, participants might consistently provide the same response to *dishwasher* as to *refrigerator* because their similarity as electrical appliances is recognized, Verheyen and Storms suggested. Similarly, the answers towards the items *psychology* and *sociology* with respect to SCIENCES might be dependent because both are social sciences.

Several other researchers have argued that consecutive categorization decisions are not made independently of one another. In a series of studies by Brooks and colleagues participants were provided with a cognitive rule, which they had to use explicitly when categorizing. Rule-irrelevant item-item similarities were found to contribute significantly to consecutive categorization judgments (Brooks & Allen, 1991; Brooks & Regehr, 1993). Similarly, Brooks, Norman, and Allen (1991) found that judgments of expert dermatologists, who are assumed to have many complex rules at their disposal for making a diagnosis, were also influenced by the similarity that existed between successively presented cases. Jones and Sieck (2003) found that commonalities between consecutive cues in a probability learning paradigm constituted sequential dependency learning, again indicating a relationship between item-item similarity and response dependencies in categorization.

The commonality between the aforementioned studies at interest is that categorization responses are influenced by similarities between the items. Indeed, intuitively one might expect that whenever a certain item resembles another for whatever reason, these items would evoke more similar categorization responses than items that do not resemble each other as much. If such a categorization pattern were to occur for items that are matched in terms of item-category similarity, this would constitute a violation of the local independence assumption. Indeed, it would suggest that item-category similarity is not the sole variable driving categorization decisions, but that item-item similarity is also of importance. Since so far only a formal argument has been given for these kinds of violations (Verheyen & Storms, 2011), we dedicate the following study to an empirical demonstration of violations of the local independence assumption.

Rationale of Studies

For the design of our study we took inspiration from a study by Hampton and Yeh (described in Hampton, 2006). For several categories we constructed quadruples of candidate exemplars. We matched the four items within a quadruple in terms of item-category similarity. In addition, each quadruple was designed to consist of two pairs of similar items, such as *rollerskates*-*skateboard* and *horse*-*mule* for the category of VEHICLES. By re-pairing the items in a quadruple one can thus construct four pairs of dissimilar items: *rollerskates*-*horse*, *rollerskates*-*mule*, *skateboard*-*horse*, and *skateboard-mule*.

For each of the six pairs of items within a quadruple one can compute the degree of categorization agreement between the items that make up a pair. If the assessment of item-category similarity is all there is to categorization, the measure of agreement for a similar pair should not differ from the measure of agreement for a dissimilar pair. After all, in constructing each quadruple it was ensured that the items were matched in terms of item-category similarity. Thus the six pairs that constitute a quadruple should not be qualitatively different if the local independence assumption holds.

This is exploited in a randomization test (Edgington & Onghena, 2007). Each of our studies will be comprised of 10 quadruples. That is, in each study 10x4 items will be presented for categorization. For each quadruple the average measure of categorization agreement across the four dissimilar pairs will be subtracted from the average measure of agreement across the two similar pairs. A test statistic is then constructed for the entire study by taking the average of these differences across the 10 quadruples that make up the study.

Figure 1 is meant to convey this procedure. The table on the left has 10 rows (one for each quadruple in a study) and 6 columns (one for each pair in a quadruple). The entries in the table are measures of categorization agreement for the items that make up a pair. In Figure 1 the row (quadruple) and column (pair) indices are substituted for the actual measures to facilitate understanding of the randomization test procedure. It is here assumed that columns 1 and 2 contain the agreement measures for the similar pairs within the quadruple (gray entries). Columns 3-6 contain the agreement measures for the dissimilar pairs within the quadruple (white entries). Dotted ellipses represent the average of the first two entries within a row. Solid ellipses represent the average of the last four entries within a row. The difference between the dotted and solid ellipses is averaged across the 10 rows to yield the test statistic.

This test statistic is then situated within a distribution of values that are arrived at by calculating the test statistic on a large number of related 10 by 6 data tables. Each of these data tables is arrived at by randomly permuting the entries within a row. This permutation implements the hypothesis that there is nothing special about the similar pairs vs. dissimilar pairs distinction. The table on the right in Figure 1 represents one such (random) re-organization of the original data. Comparing the values in the first two columns of the table with those in the last four columns of the table no longer entails a comparison of pairs comprised of similar items and pairs comprised of dissimilar items (gray and

(1,1)	(1,2)	(1,3)	(1,4)	(1,5)	(1,6)
(2,1)	(2,2)	(2,3)	(2,4)	(2,5)	(2,6)
(3,1)	(3,2)	(3,3)	(3,4)	(3,5)	(3,6)
(4,1)	(4,2)	(4,3)	(4,4)	(4,5)	(4,6)
(5,1)	(5,2)	(5,3)	(5,4)	(5,5)	(5,6)
(6,1)	(6,2)	(6,3)	(6,4)	(6,5)	(6,6)
(7,1)	(7,2)	(7,3)	(7,4)	(7,5)	(7,6)
(8,1)	(8,2)	(8,3)	(8,4)	(8,5)	(8,6)
(9,1)	(9,2)	(9,3)	(9,4)	(9,5)	(9,6)
(10,1)	(10,2)	(10,3)	(10,4)	(10,5)	(10,6)

(1,6)	(1,3)	(1,5)	(1,1)	(1,2)	(1,4)
(2,5)	(2,1)	(2,2)	(2,3)	(2,4)	(2,6)
(3,4)	(3,5)	(3,2)	(3,3)	(3,6)	(3,1)
(4,4)	(4,3)	(4,1)	(4,5)	(4,6)	(4,2)
(5,6)	(5,4)	(5,5)	(5,1)	(5,3)	(5,2)
(6,2)	(6,4)	(6,5)	(6,3)	(6,1)	(6,6)
(7,4)	(7,2)	(7,6)	(7,5)	(7,1)	(7,3)
(8,3)	(8,5)	(8,4)	(8,6)	(8,1)	(8,2)
(9,6)	(9,3)	(9,5)	(9,4)	(9,1)	(9,2)
(10,1)	(10,6)	(10,4)	(10,2)	(10,5)	(10,3)

Figure 1: Rationale of the randomization test used to detect violations of local independence.

white entries are dispersed over the columns). If the value of the test statistic for the original data is located in the tail of the resulting distribution, this is an indication that, contrary to the local independence assumption, there is something peculiar about the distinction.

Pilot

We started our project with an exploratory study, which was intended to obtain typicality norms for 1276 items from 24 categories. For practical reasons we split up the materials in two subsets. Set 1 comprised 616 items from 12 categories (CLOTHING, DISEASES, ENERGY SOURCES, FISH, FRUITS, KITCHEN UTENSILS, MUSICAL INSTRUMENTS, NATURAL EARTH FORMATIONS, PROFESSIONS, SCIENCES, VEHICLES, WEAPONS). Set 2 comprised 660 items from 12 different categories (ADDICTIONS, ANIMALS, BUILDINGS, CRIMES, FURNITURE, INSECTS, MEDIA, SPORTS, TOOLS, TOYS, VEGETABLES, WEATHER PHENOMENA)[1]. We primarily focused on borderline items such as *rollerskates* for VEHICLES or *bullfighting* for SPORTS: items that are neither evident members, nor evident non-members of the target categories. However, several clear members and non-members were included for each category to make the task less transparent and more rewarding for the participants.

We had 29 first year Bachelor of Psychology students provide typicality judgments on a seven point Likert scale ranging from *very atypical* to *very typical*. Fifteen students provided judgments for Set 1. Fourteen students provided judgments for Set 2. Every participant was presented with a different order of categories and items within categories. The reliability of the typicality judgments ranged from .87 for NATURAL EARTH FORMATIONS to .98 for ANIMALS with a mean of .94.

We continued by selecting from the materials in the exploratory study 20 quadruples, 10 belonging to Set 1 and 10 belonging to Set 2. The four items that were selected for each quadruple were to adhere to the following criteria: (i) The four items in a quadruple had to match in terms of mean typicality. (ii) The items in a quadruple were to fall apart in two pairs of semantically similar items. (iii) The items had to be borderline items, meaning they displayed intermediate typicality ratings.

Mean typicality is often used as a proxy for item-category similarity (Hampton, 1998, 2007; Verheyen, Hampton, & Storms, 2010). Criterion (i) thus ensures that the items within a quadruple are matched in terms of item-category similarity, the latent variable put forward to explain the range of semantic categorization data.

Several studies have identified item-item similarity as a factor affecting categorization (Brooks & Allen, 1991; Brooks, Norman, & Allen, 1991; Brooks & Regehr, 1993; Jones & Sieck, 2003; Verheyen & Storms, 2011). Criterion (ii) was included to look for violations of local independence due to item-item similarity. Such violations occur when categorization agreement is found to be higher between items in a semantically similar pair than between items in a semantically dissimilar pair.

Categorization agreement between respondents is maximal at the high and low ends of the typicality scale. Items of intermediate typicality, on the other hand, elicit inter-individual categorization differences (Hampton, 1998, 2007; Verheyen, Hampton, & Storms, 2010). Criterion (iii) was incorporated because we planned to look at categorization agreement differences between similar and dissimilar pairs to detect violations of local independence. This requires investigating items for which categorization agreement differences are likely to occur.

Criterion (iii) was met by the inclusion of items with a mean typicality score of around 4 (i.e., the middle of the seven point typicality scale we used; *M*=3.4, *SD*=1.3). Criterion (ii) was met by verifying whether the participants in the Pilot Study gave similar responses to the items within a pair considered to be similar, and dissimilar responses to the items within a pair considered to be dissimilar. This was done by calculating across respondents the correlation between the typicality judgments for the items in a quadruple (see also Hampton & Yeh, described in Hampton, 2006). This results in 6 correlations for the 6 item pairs in a quadruple. We required the correlations for the similar pairs to exceed 0.3 (*M*=.76, *SD*=.36), and the correlations for the dissimilar pairs to be lower than the lowest of the similar pair correlations (*M*=.24, *SD*=.26).

To determine whether Criterion (i) was met, we conducted a permutation test (Edgington & Onghena, 2007). For each of the six pairs in a quadruple we compared the mean difference in typicality rating across respondents with a distribution of values that resulted from computing the same difference under the assumption that there was no difference between the ratings for the items comprising a pair. To this end we randomly determined whether a participant's ratings for the items comprising a pair were to be swopped or not. If the items were truly matched in terms of typicality this procedure shouldn't result in mean differences in typicality across participants that are different from the one observed for the empirical (unswopped) data. We concluded that a pair of items was matched in terms of mean typicality if the empirical mean typicality difference fell between the 2.5th percentile and the 97.5th percentile of the distribution comprising 10,000 reference values simulated under the assumption of no difference. Both for the Set 1 quadruples and for the Set 2 quadruples this was the case for 57 of the 10 x 6 pairs. The three violations in Set 1 were due to FRUIT item pairs. The violations in Set 2 were due to the categories ADDICTIONS (1 pair) and TOYS (2 pairs)[2].

Space restrictions do not allow us to provide all the details about the selected materials. We can, however, provide an example to illustrate the employed procedure. The pairs *rollerskates-skateboard* and *horse-mule* were considered candidates to form a quadruple for VEHICLES. The first two

[1]Note that in all studies items and categories were presented in Dutch. We provide their English translation for convenience.

[2]These three quadruples were nevertheless presented to participants in Studies 1 and 2. None of the results we present depend on whether they are included in the analyses or not.

items may make up a semantically related pair because both are means of transportation that one could buy in a sports or toy store. They can be contrasted with the two last items, which are both animals that can be used for transportation. With mean typicality scores of 4.50, 4.57, 4.21, and 4.71 they can be considered borderline items. The correlations between the typicality ratings for the within-pair items are .86 and .82. These are considerably higher than the values for the across-pairs correlations (.54, .65, .66, .57) indicating higher item-item similarity within pairs than across pairs. The permutation test indicated that in all six pairs that can be formed with the four items, the items were matched in terms of mean typicality. Having met Criteria (i)-(iii), these items were presented for categorization in Studies 1 and 2, along with other items such as *sink-toilet* and *umbrella stand-coat rack* for FURNITURE and *plastic bag-pillow* and *pest control-weed killer* for WEAPONS.

Note that for some categories it was possible to determine several such quadruples, but we chose to include only one per category in our studies. This way we could be fairly certain that any dependencies in categorization were to be attributed to the particular higher-order similarity structure of the quadruple, and were not due to relationships with other items that were presented.

Study 1

In this study we presented the 10 quadruples that were selected from Set 1 to a group of 120 first year Bachelor of Psychology students. We presented the 10 quadruples that were selected from Set 2 to a different group of 124 first year Bachelor of Psychology students. None of the Study 1 participants had participated in the Pilot. For every participant we randomized the order in which the 40 item-category combinations were presented. We asked each participant to provide binary membership judgments. Participants could indicate they felt an item was a member of the target category by circling a 1, indicate they felt an item was not a category member by circling a 0, or indicate that they did not know the item and/or category.

To analyze the data we employed the permutation test that was outlined in section *Rationale of Studies*. A separate analysis was conducted for the quadruples from Set 1 and those from Set 2. As a measure of categorization agreement within a pair we computed Cohen's kappa coefficient (Cohen, 1960) which yields a value between 0 and 1. A value of 0 indicates that any agreement on behalf of the participants concerning whether two items should be considered category members or not is purely coincidental. A value of 1 indicates that the participants are in perfect agreement concerning category membership. One can think of values between 0 and 1 as reflecting the "distance" travelled from coincidental agreement to perfect agreement. A data table like the one on the left in Figure 1 can be constructed by computing the kappa coefficient for all 10x6 item pairs. The row for the category VEHICLES, for instance, would have values .51 and .65 in columns 1 and 2 for the within-pairs categorization agreement. It would have values .31, .24, .31, .23 in columns 3-6 for the across-pairs categorization agreement. Clearly, taking the difference between the mean of the former two values and the mean of the latter four values results in a positive value, suggesting higher categorization agreement within pairs than across pairs. That is, there was a tendency among participants either to indicate that *rollerskates* and *skateboards* are VEHICLES, but *mules* and *horses* are not, or to indicate that *mules* and *horses* are VEHICLES, but *rollerskates* and *skateboards* are not. This trend is apparent across all quadruples from a set, as is evidenced by the positive value for the empirical test statistic that is arrived at by computing the average of the mean kappa differences across quadruples (dashed vertical line in both panels of Figure 2).

The empirical test statistic was compared with 100,000 simulated test statistics, resulting from randomly permuting the kappa values for a quadruple. (Permutation in one quadruple was independent of the permutation in another quadruple.) This procedure effectively implements the hypothesis that there is nothing peculiar about the distinction between similar pairs and dissimilar pairs. The

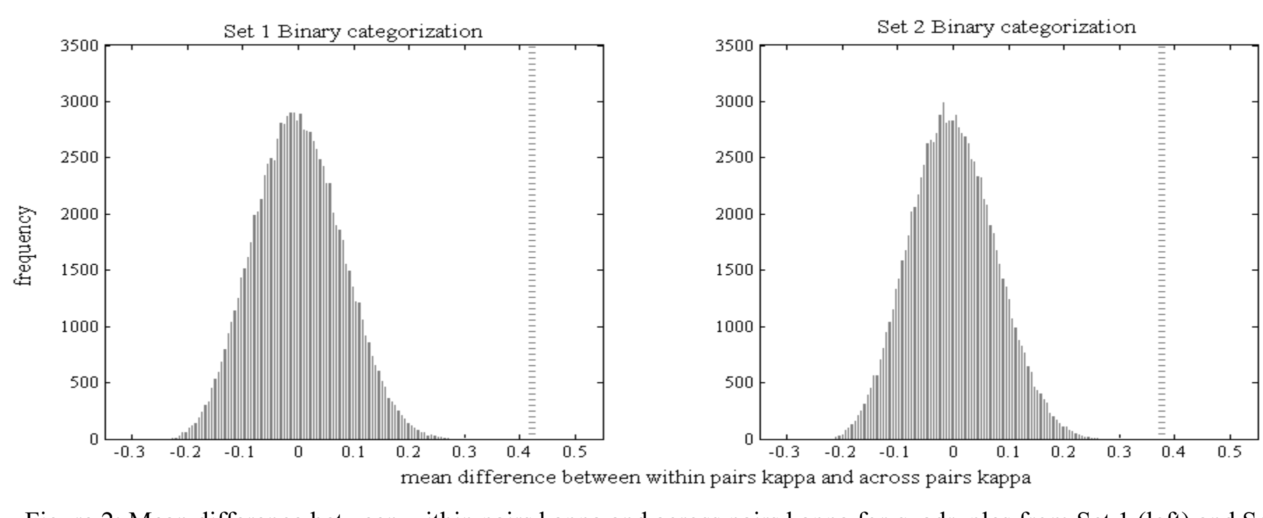

Figure 2: Mean difference between within pairs kappa and across pairs kappa for quadruples from Set 1 (left) and Set 2 (right) in Study 1. Observed (dashed vertical line) vs. simulated difference under local independence assumption (histogram).

resulting distribution is depicted by the histograms in Figure 2. They reflect the range of categorization agreement differences one would expect if local independence were to hold. The empirical test statistic is clearly greater than what one would expect under the local independence assumption. It is located in the upper 5% of the distribution. The nature of the quadruple manipulation suggests that item-item similarity is likely to be responsible for the violations.

Study 2

The purpose of Study 2 was to replicate the results of Study 1 using a different categorization measure. We presented 51 students with the 20 quadruples we obtained from our pilot study. None of the 51 participants had taken part in the previous studies or were aware of them. This time the participants were invited to respond to each of the 80 item-category combinations with a continuous membership judgment using an 11 point Likert scale ranging from *clearly does not belong to the category* to *clearly belongs to the category* (following the instructions used in Estes, 2004). We presented the quadruples from Set 1 and Set 2 in two distinct blocks separated by a short break and alternated between participants which block was presented first. Within each block we also randomized the presentation order of the 40 item-category combinations. The presentation of items and the acquisition of continuous categorization responses were computerized. On each trial a single item-category combination was displayed. After each response the screen was cleared and a new item-category combination was presented. We adopted this computerized procedure to exclude the possibility that the interdependency effect in Study 1 was due to participants making their later responses consistent with previous ones, by backtracking answers when they recognized some relationship between two items. (Study 1 was a paper and pencil task.)

To analyze the data we again employed the permutation test that was outlined above. As was the case for Study 1 a separate analysis was undertaken for the quadruples from Set 1 and those from Set 2. As a measure of categorization agreement for two items we computed the correlation between the continuous categorization ratings given to each of the items, across the different participants in the study. For example, the correlation was calculated across participants of the ratings given to each member of a pair of similar items such as *horse-mule* (.64) and of the ratings given to each member of a dissimilar pair such as *rollerskate-mule* (.25). For the computation of the empirical and simulated test statistics we applied the Fisher transformation to these correlations. The number of simulations and the criterion were again set to 100,000 and .05, respectively.

In Figure 3 the empirical test statistic is again depicted by dashed vertical lines. Both for the Set 1 quadruples (left) and the Set 2 quadruples (right), the categorization judgments appear more in line for items that form similar pairs than for items that form dissimilar pairs. As was the case for Study 1, the empirical test statistic is located in the upper 5% of the distribution of test statistics that was arrived at by assuming local independence. Not only do these results demonstrate violations of the independence assumption; they suggest that these violations might be the result of item-item similarity. The results from Study 2 add to those from Study 1 in that they show the interdependencies to hold, regardless of whether categorization judgments are made in a binary or in a continuous fashion.

Discussion

Like most accounts of semantic categorization the Threshold Theory assumes that similarity drives categorization decisions (Hampton, 1998, 2007). Verheyen, Hampton, & Storms (2010) have provided a formalization of the Threshold Theory that explains semantic categorization in terms of a latent variable. They showed that this latent variable was strongly related to mean typicality, a measure that is considered to be a proxy of item-category similarity. Local independence is one of the

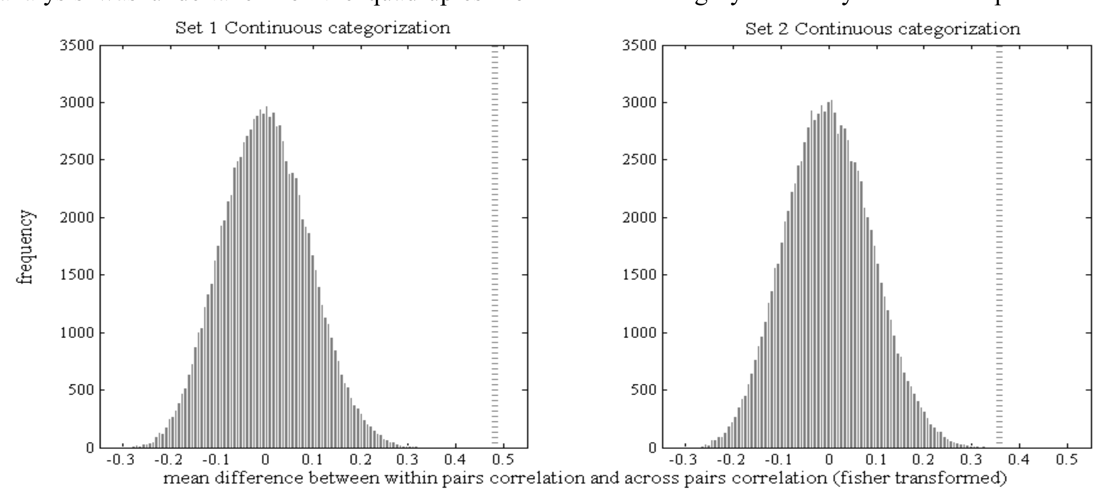

Figure 3: Mean difference between within pairs correlation and across pairs correlation for quadruples from Set 1 (left) and Set 2 (right) in Study 2. Observed (dashed line) vs. simulated difference under local independence assumption (histogram).

underlying assumptions of latent variable models. Categorization decisions towards different items should be conditionally independent of each other given the items' scores on the latent variable (i.e, their mean typicality or item-category similarity). Verheyen and Storms (2011) already demonstrated that the local independence assumption was not tenable by establishing that the application of the latent variable model to semantic categorization data left some of the structure in the data unexplained. In addition to this formal argument the current studies offer an empirical demonstration of violations of the local independence assumption.

The current results need not be taken to discard similarity-based theories of categorization in general or the Threshold Theory in particular. To the contrary, the current results clearly establish that participants' decisions are subject both to item-category similarity and item-item similarity. The former shows in the inter-individual categorization differences for the items of intermediate typicality that comprise our materials. The latter shows in that participants' categorization decisions are in line with the 2 x 2 similarity structure of the quadruple design. Similarity thus clearly emerges as an important explanatory construct in categorization research.

We see at least two manners in which categorization theories can be extended to account for the current findings. One is to recognize that individuals may weight category information differently (because of previous experiences and immediate situations; e.g., Smith & Samuelson, 1997), resulting in different assessments of item-category similarity. In fact, the typicality correlation procedure we used to validate Criterion (ii) of our stimulus selection procedure might indicate just that: some individuals might attach weight to aspects of the VEHICLES category that make *rollerskates* and *skateboards* more typical exemplars than *horses* and *mules*, while others might attach weight to aspects that make *horses* and *mules* more typical VEHICLES than *rollerskates* and *skateboards.* In the words of Hampton (2006, p. 99): "*Having the same feature profile, similar pairs would move up or down together as feature weights change across individual raters, whereas dissimilar pairs would not*". Verheyen and Storms (2013) have implemented a mixture version of the latent variable Threshold model that identifies latent subgroups of participants who weight features of the category differently, resulting in a different organization of the items along the latent variable. Alternatively, one could explicitly acknowledge the dependencies by incorporating them in the modeling framework. This could amount to bringing additional similarity information into the model (item-item similarity in addition to the already present item-category similarity). This possibility is considered by Verheyen and Storms (2011).

Acknowledgments

Jesse Zee completed Studies 1 and 2 as part of a research internship while pursuing the degree of Bachelor of Psychology at the University of Leuven. Steven Verheyen is a postdoctoral fellow at the Research Foundation-Flanders. We would like to thank James A. Hampton for his helpful comments throughout various stages of the project.

References

Allen, S. W., & Brooks, L. R. (1991). Specializing the operation of an explicit rule. *Journal of Experimental Psychology: General*, *120*, 3-19.

Brooks, L. R., Norman, G. R., & Allen, S. W. (1991). Role of specific similarity in a medical diagnostic test. *Journal of Experimental Psychology: General, 120*, 278-287.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20*, 37-46.

Edgington, E., & Onghena, P. (2007). *Randomization tests*. Boca Raton: Chapman & Hall/CRC.

Estes, Z. (2004). Confidence and gradedness in semantic categorization: Definitely somewhat artifactual, maybe absolutely natural. *Psychonomic Bulletin & Review, 11*, 1041-1047.

Hampton, J. A. (1998). Similarity-based categorization and fuzziness of natural categories. *Cognition, 65*, 137-165.

Hampton, J. A. (2006). Concepts as prototypes. *The Psychology of Learning and Motivation: Advances in Research and Theory, 46*, 79-113.

Hampton, J. A. (2007). Typicality, graded membership, and vagueness. *Cognitive Science, 31*, 355-384.

Jones, M., & Sieck, W. R. (2003). Learning myopia: An adaptive recency effect in category learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 29*, 626-640.

Koendrink, J. (1999). Guest editorial. *Perception, 28,* 1181- 1184.

Regehr, G., & Brooks, L. R. (1993). Perceptual manifestations of an analytic structure: The priority of holistic individuation. *Journal of Experimental Psychology: General, 122,* 92-114.

Smith, L. B., & Samuelson, L. K. (1997). Perceiving and remembering: Category stability, variability and development. In K. Lamberts, & D. Shanks (Eds.), *Knowledge, concepts, and categories* (pp. 161-195). East Sussex, UK: Psychology Press.

Verheyen, S., Hampton, J. A., & Storms, G. (2010). A probabilistic threshold model: Analyzing semantic categorization data with the Rasch model. *Acta Psychologica, 135*, 216-225.

Verheyen, S., & Storms, G. (2011). Towards a categorization-based model of similarity. In L. Carlson, C. Hölscher, & T. F. Shipley (Eds.), *Proceedings of the 33rd Annual Conference of the Cognitive Science Society* (pp. 614-619). Austin, TX: Cognitive Science Society.

Verheyen, S., & Storms, G. (2013). A mixture approach to vagueness and ambiguity. *PloS ONE, 8*: e63507