Towards a Categorization-Based Model of Similarity

Steven Verheyen (steven.verheyen@psy.kuleuven.be)
Gert Storms (gert.storms@psy.kuleuven.be)
Department of Psychology, University of Leuven
Tiensestraat 102, B-3000 Leuven, Belgium

Abstract

Most accounts of categorization assume the categorization decision for an item to be independent of the categorization decisions for other items. A number of methods are brought to bear on the question of whether this assumption is justified. These methods involve the application of a formal categorization model that explicitly incorporates the independence assumption to categorization data and the subsequent investigation of the residuals for unexplained structure. The residuals reveal multiple departures from independence, suggesting that the independence assumption in many a categorization account should be relaxed. Following this suggestion the applied formal model is extended to allow for dependent categorization decisions. It is explained how the extended model might address the concern that categorization accounts have erred in using similarity as an explanatory construct. It promises to be a significant step towards a categorization-based model of similarity.

Keywords: categorization; similarity; threshold theory.

Introduction

Similarity is arguably the variable that is most often invoked to explain categorization decisions. According to most accounts of categorization an item is believed to be a category member if its representation sufficiently resembles the category representation (regardless of whether the latter is believed to be an abstracted summary representation, an instantiated set of representative exemplars, an ideal, or a coherent theory). The Threshold Theory of categorization, for instance, posits that prior to making a categorization decision the similarity between the item's representation and the category's representation is compared against an internal threshold (Hampton, 2007). If the assessed similarity exceeds the threshold, the item will be endorsed as a category member; otherwise it will not.

Most accounts of categorization will make the additional assumption that consecutive categorization decisions are made independently from one another. It is believed that every new item that is encountered for categorization will invoke the same similarity-assessment procedure that earlier items have. That is, participants will provide a categorization decision by determining whether the new item's representation sufficiently resembles the category's representation. The answer that is provided on this particular categorization trial is thus believed to be arrived at independently from the decisions that were made earlier (and the ones that await). In the framework of the Threshold Theory, for instance, every new categorization decision entails a comparison of the item-category similarity against the internal threshold, without regard of the decisions for alternate items.

This assumption of independence might prove too strong, particularly in the context of the tasks that are employed to study categorization in natural language categories. The list of items that is presented for categorization is generally a mix of clear members, borderline members, and clear nonmembers of the target category. However, even the clear nonmembers of the target category in these tasks are chosen to be at least somewhat related to the category's most prototypical items. For instance, the nonmembers included for categorization in a category like ***vegetables*** tend to be comprised of other food or plant items. Employing items from the animals or artifacts domains instead, would presumably render the task less ecological valid and (perhaps more to the point) might detract from the similarity-based processes we intend to study with these tasks. Importantly, when all the items that make up a set of potential category members are drawn from a single domain (be it animals, artifacts, foods, activities, ...) it is likely that meaningful similarity relations exist among them. These similarity relations might impose structure on the corresponding categorization decisions that have been neglected in treating these decisions as independent from one another. For instance, in categorizing items as ***vegetables*** or not, one can imagine participants consistently giving the same response to *parsley* as to *sage* when their similarity as ***herbs*** is recognized.

In what follows we will employ various methods to establish that the data that result from the traditional categorization task violate the assumption of independence. These methods originate from the item response models literature, where departures from independence are known as Local Item Dependencies (LIDs). Rather than considering LIDs as nuisances that one is better off eliminating (which is common practice in the item response models literature), we will relate the LIDs to ratings of item-item similarity to argue that they are a substantial part of categorization decisions which future accounts of categorization will have to incorporate. The case for the existence and importance of LIDs in categorization will be made by means of a reanalysis of previously published categorization data using the Rasch model (Rasch, 1960). The reasons for using the Rasch model to introduce one of the shortcomings of many current categorization accounts are threefold. (i) The model naturally accounts for the inter-individual differences in categorization that are characteristic of the natural language categories we study (Verheyen, Hampton, & Storms, 2010). (ii) Since the Rasch model is an item response model it is straightforward to apply existing methods for detecting LIDs to it. (iii) A Rasch-like model that accommodates the need to incorporate LIDs offers the intriguing possibility of deriving similarity from categorization, instead of the other way around.

The next section provides additional information about the Rasch model. In our exposition of the model we will use indices c to refer to categorizers (n_c in total) and indices i to refer to items (n_i in total).

Applying the Rasch Model to Categorization

In the Rasch model individual categorization decisions x_{ci} are considered the outcomes of Bernoulli trials with probability p_{ci}. Equation (1) expresses how p_{ci} is fully determined by the values of θ_c and β_i. It expresses that the more β_i exceeds θ_c on a latent scale, the higher the probability is that categorizer c will endorse item i, and vice versa.

$$p_{ci} = \frac{e^{(\beta_i - \theta_c)}}{1 + e^{(\beta_i - \theta_c)}} \tag{1}$$

This formalization is reminiscent of the Threshold Theory's claim that categorization decisions arise from the assessment of the similarity between the item's representation and the category's representation (Hampton, 2007). This assessment results in the positioning of the item along a latent similarity scale (i.e., fixing the item's β_i value). The further along the scale an item is positioned, the higher its similarity to the category is assumed to be. According to the Threshold Theory categorizers then impose threshold criteria on the scale to determine whether the assessed similarity affords a positive rather than a negative categorization decision. The value of θ_c is taken to indicate the position of a categorizer's threshold criterion. The probability expressed in Equation (1) decreases with θ. Low values of θ_c indicate rather liberal categorizers for whom a modest degree of similarity suffices to conclude category membership. High values of θ_c characterize more conservative categorizers who require extensive similarity between item and category to conclude category membership. Differences in the estimates of θ_c allow the Rasch model to account for the variable extension of natural language categories.

Indeed, Verheyen et al. (2010) applied the model to 250 participants' categorization decisions towards potential exemplars of eight natural language categories to show that it accords with the inter-individual differences in the data. Moreover, they also found the model to reconcile the counterintuitive finding that the items afford both binary membership decisions and continuous typicality ratings. The estimates of the items' positions along the latent scale correlated almost perfectly with independently provided ratings of their typicality. This accords with the Threshold Theory's assertion that there exists a linear relationship between item-category similarity and typicality.

In the following section we elaborate on the Verheyen et al. (2010) categorization data that we will reanalyze to look for LIDs. To evaluate whether these LIDs can be interpreted in terms of item-item similarity, an external measure of similarity is required. We also discuss the procedure used to obtain such a measure.

Data

Categorization

Two hundred and fifty first year psychology students at the University of Leuven completed a categorization task for partial fulfillment of a course requirement. The task included 8 categories with 24 items each. The categories consisted of two animal categories (***fish*** and ***insects***), two artifact categories (***furniture*** and ***tools***), two food categories (***fruits*** and ***vegetables***), and two activity categories (***sciences*** and ***sports***). The corresponding category items included clear members, clear nonmembers, and borderline cases. The data collection took place in a large classroom where all participants were present at the same time. Each of them was handed an eight-page questionnaire to fill out. They were told to carefully read through the 24 items on each page and to decide for each item whether or not it belonged in the category printed on top of the page. Participants indicated their answer by either circling 1 for *yes* or 0 for *no*. Five different orders of category administration were combined with 2 different orders of item administration, resulting in 10 different questionnaires. Each of these was filled out by 25 participants.

Similarity

Ninety-two first year psychology students at the University of Leuven provided pairwise similarities for partial fulfillment of a course requirement. Eighty-three of them provided ratings for the 276 item pairs of one of the eight categories of ***fish***, ***insects***, ***furniture***, ***tools***, ***fruits***, ***vegetables***, ***sciences***, and ***sports***. Six students provided ratings for the pairs of two of these categories, two students provided ratings for three categories' pairs, and one student rated the item pairs of seven categories. All the items that belonged to a particular set were shown before the onset of the similarity rating task. This procedure served two purposes: We wanted to ensure that participants had an idea about the degree of similarity and difference that was represented in the set and we wanted them to only rate the pairs from a set of which they knew all the items. Because participants did not always know all the items in a particular set, they sometimes had to complete a different set than was originally intended. Because of this not all sets were completed by the same number of participants. For ***fish***, ***insects***, ***furniture***, ***tools***, ***fruits***, ***vegetables***, ***sciences***, and ***sports*** pairwise similarities were provided by 16, 15, 14, 15, 15, 12, 10, and 11 participants, respectively. Each of them rated the similarity of every item pair by providing a number between 1 (*totally dissimilar*) and 20 (*totally similar*). Item pairs were presented in a random order. The presentation order of items within a pair was also randomized. The ratings were averaged across participants and reliabilities were estimated by split-half correlations corrected with the Spearman-Brown formula. The reliability estimates ranged between .86 for ***vegetables*** and .93 for ***sciences***.

```
model {
  #Endorsing An Item Is A Bernoulli Trial
  for (c in 1:nc){
      for (i in 1:ni){
           x[c,i] ~ dbern(p[c,i])
      }
  }
  #Probability Is Determined By Categorizer-Item Positions
  for (c in 1:nc){
      for (i in 1:ni){
           p[c,i] <- (exp(beta[i]-theta[c])/(1+exp(beta[i]-theta[c])))
      }
  }
  #Priors For Categorizers And Items
  for (c in 1:nc) { theta[c] ~ dnorm(0,clambda) }
  clambda ~ dgamma(.001,.001)
  csigma <- 1/sqrt(clambda)
  for (i in 1:ni) { beta[i] ~ dnorm(0,.001) }
}
```

Figure 1: WinBUGS code for the Rasch model.

Detecting Local Item Dependencies

The Rasch model assumes that categorization decisions are made independently from one another. It not only shares this assumption with many competing categorization models, but also with the majority of other item response models. As a consequence, many means of detecting violations of independence have been put forward in the item response models literature. In this section we will employ two of these to investigate whether such violations also characterize categorization data. The subsequent section will be devoted to the substantive interpretation of these results. In order to detect LIDs we first need to apply the Rasch model to the categorization data that were described earlier. This entails a replication of the modeling exercise conducted by Verheyen et al. (2010).

We implemented the Rasch model in WinBUGS (Lunn, Thomas, Best, & Spiegelhalter, 2000) using the code that is provided in Figure 1. Note that a normal distribution is defined over the θ_c's. We opted to estimate the standard deviation of this distribution for every category, instead of including a separate scaling parameter α for every category as was done in Verheyen et al. (2010). We employ the posterior means across 5 chains with 10,000 samples each as point estimates for the various model parameters.

Graphical Model Evaluation

One way of assessing whether the categorization data present with LIDs is graphical in nature and incorporates the rational behind parametric bootstrapping (Tuerlinckx & De Boeck, 2004). Figure 2 includes the results of applying this procedure to the categorization data of the ***sciences*** category. For all possible item pairs in the category, the empirical log odds ratio was computed:

$$log\Big(\frac{(n_{11}+.5)(n_{00}+.5)}{(n_{10}+.5)(n_{01}+.5)}\Big) \qquad (2)$$

and converted in a gray-scale value. (In Equation (2) n_{11} is defined as the observed frequency of a joint $(1,1)$-response for the item pair under study, and so on.) These values were subsequently placed in the upper triangular part of the left most square matrix in Figure 2. The lighter a square in the matrix, the higher the value for the corresponding empirical log odds ratio is. If the Rasch model is able to capture the dependencies the categorization data present with, the square matrix on the left should not differ greatly from the four square matrices on the right. They represent the log odds ratios obtained from four simulated data sets. Each of these data sets was constructed using the parameter estimates that resulted from fitting the Rasch model to the empirical categorization data. It would appear that the matrix holding the empirical log odds ratios contains somewhat more structure than the matrices holding the simulated log odds ratios do. The results in Figure 2 would thus suggest that there are dependencies present in the categorization data that the Rasch model doesn't capture. The same conclusion follows inspection of similar figures for the remaining categories.

Although the graphical model evaluation procedure has the benefit that it provides an overview of discrepancies pertaining to dependencies at a glance, the extent of the discrepancies remains subject to interpretation. One could of course look deeper into the matrices to find out for which item pairs the greatest differences emerge, but no principled means of deciding which of these differences constitute substantial ones is readily available. The following method for detecting LIDs provides a manner to remedy this.

Correlations between Standardized Residuals

Given the parameter estimates of an applied model, evidence of local dependence between two items can also be obtained by calculating the correlation between the residuals of the observed and expected responses (Andrich & Kreiner, 2010). This correlation provides a window into the association between items that is left unexplained by the model with its assumption of independence. Following the estimation of the Rasch parameters we thus obtained the correlation between the standardized residuals for each pair of items in a category. Wright (1977) provides the following formula for the computation of individual standardized residuals:

$$RES_{ci} = \frac{x_{ci} - p_{ci}}{\sqrt{w_{ci}}} \text{ with } w_{ci} = p_{ci} \times (1 - p_{ci}) \qquad (3)$$

In a regular search for LIDs the item pairs with the highest correlations between their residuals are identified as violating the independence assumption. It is common practice to set a .05 significance criterion for this (e.g., Andrich & Kreiner, 2010). If we were to employ this heuristic 4 item pairs from the ***fish*** category would be identified as locally dependent. For the categories of ***insects***, ***furniture***, ***tools***, ***fruits***, ***vegetables***, ***sciences***, and ***sports***, these numbers equal 5, 5, 2, 1, 4, 2, and 5, respectively. Together they constitute evidence that violations of the independence assumption exist in the categorization data. The results add to those of the graphical model evaluation procedure in that they are explicit about the item pairs one should sit up and take notice of. Table 1 holds for every category an example of an item pair with significantly correlated residuals. The nature of these items suggests that item-item similarity might be responsible for their

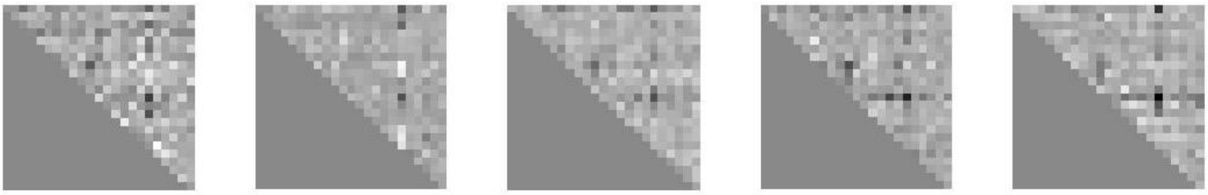

Figure 2: Log odds ratio plot of the empirical categorization data (first matrix) and four data sets simulated from the Rasch model estimates (matrices two to five) for the category of ***sciences***.

Table 1: Examples of item pairs that violate independence.

Category	Items	
fish	crab	lobster
insects	mosquito	wasp
furniture	dishwasher	refrigerator
tools	pitchfork	rake
fruits	acorn	pine cone
vegetables	lettuce	spinach
sciences	psychology	sociology
sports	billiards	darts

dependence. 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. The answers towards the items *psychology* and *sociology* with respect to ***sciences*** might be dependent because both are ***social sciences***. Similar arguments can be constructed for the other item pairs in Table 1.

Understanding Local Item Dependencies

Correlation Analyses

Identifying some item pairs as dependent (and implicitly all others as independent) by means of significance tests has a certain amount of arbitrariness to it. After all, the pattern of correlations between residuals might carry more information than is revealed by merely separating the significant ones from the insignificant ones. Alternatively, one can make use of all the dependency estimates. These can then be related to the ratings of item-item similarity we collected.

As a first approximation of the relationship between the dependency estimates and the item-item similarities their Pearson correlation was calculated. The correlation was established at .43 for ***fish***, at .57 for ***insects***, at .34 for ***furniture***, at .29 for ***tools***, at .30 for ***fruits***, at .35 for ***vegetables***, at .54 for ***sciences***, and at .26 for ***sports***. All of these correlations are significant at the .0001 level of significance (according to one-tailed t-tests).

Of course, the significance of a correlation of .26 should not be overstated in light of the large number of observations ($N = 276$). The correlation between the dependency estimates and item-item similarity is by no means perfect. However, one does need to take into consideration that these correlations with item-item similarity were obtained after typicality was partialled out. Indeed, application of the Rasch model to the categorization data in Verheyen et al. (2010) yielded β_i estimates that correlated between .94 and .98 with rated typicality. As a consequence it is safe to say that the information conveyed by the residuals is typicality-free. Provided that typicality is one of the major organizing principles of similarity ratings that have been obtained in the context of a semantic category (e.g., Verheyen, Ameel, & Storms, 2007) we believe it to be substantial that reliable correlations with item-item similarity remain after the variable is partialled out. It suggests that in addition to item-category similarity (i.e., typicality) item-item similarities inform categorization decisions in natural language categories. That is to say, a categorization decision pertaining to one item isn't necessarily independent from a categorization decision pertaining to another item.

Procrustes Analyses

It might be more appropriate to scale the dependency estimates and similarity ratings than to employ the raw data. The scaled association in which a particular item features is informed by all the available information about that item (i.e., the various other associations in which the item features) and therefore considered more reliable. For instance, when multidimensional scaling (MDS) is employed to approximate associations in a geometrical space, the coordinates of the items are determined such that the distances between them optimally reflect their association (the greater the association between two items, the smaller their distance in the space). Since the distances in the space have to fulfill the triangle inequality, the positioning of two items is co-determined by the other associations in which the items feature. Using MDS to spatially represent dependencies/similarities has the added advantage that it entails the prevailing means of representation that many models of categorization subscribe to (e.g., exemplar models, Nosofsky, 1984; prototype models, Smith & Minda, 1998; ...) and many researchers have used to predict category-related behaviors (e.g., processing of analogies, Rumelhart & Abrahamsen, 1973; inductive strength, Rips, 1975, ...).

The estimated item dependencies and the rated item-item similarities for the eight categories were therefore subjected to nonmetric multidimensional scaling. Configurations were obtained in dimensionality 2 to 6[1].

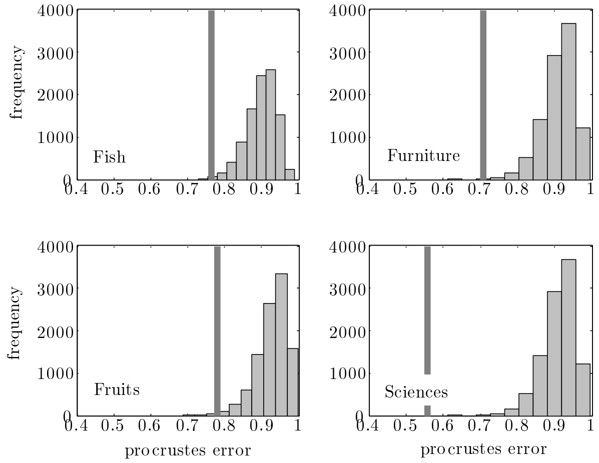

Figure 3: Comparison of empirically observed procrustes errors (vertical lines) with distributions of procrustes errors for randomly generated configurations (histograms).

Whether the MDS solutions for the dependency estimates correspond to those for the item-item similarities can be evaluated through procrustes analyses (Borg & Groenen, 2005). In a procrustes analysis two spatial representations resulting from MDS are fitted as closely to each other as possible using linear transformations. The analysis yields an error measure that reflects the degree of mismatch that remains after these transformations. For each category we employed procrustes analyses to fit each spatial configuration of the dependency estimates (#5, dimensionality 2 to 6) to each spatial configuration of the item-item similarities (#5). The 25 procrustes errors that were in this manner obtained for every category were compared against the procrustes errors that were obtained through the comparison of randomly generated spatial configurations with the same dimensionalities. Spatial configurations of items in a particular dimensionality were constructed by randomly sampling the coordinates of the items on all dimensions from a uniform distribution between 0 and 1000. Twenty-four items were included; the same as the number of items in the empirical data. This procedure was repeated $10,000$ times in each of the dimensionalities 2 to 6. These randomly generated spatial configurations were then subjected to the same procrustes analyses we employed to compare the spatial representations for the item dependencies and the item-item similarities. If the configurations for the latter two data sets carry alignable information, the resulting error from the procrustes analysis should be less than that obtained from the randomly generated configurations.

In the 200 (8 categories × 5 × 5 dimensionalities) comparisons that were conducted, the empirically observed procrustes errors fell 192 times (96%) below the 5% value of the distributions of procrustes errors generated from random configurations. This constitutes strong evidence that the commonalities that the procrustes analyses picked up are not due to coincidence. That is to say, the configurations that represent the item dependencies share structure with the configurations that represent the item-item similarities. Again, one needs to take into consideration that this correspondence emerges despite the fact that the dependency estimates are typicality-free. For four of the categories Figure 3 locates the empirically observed procrustes error in the distribution of procrustes errors from comparing random configurations. Each vertical line represents the procrustes error from comparing one dependency configuration with one similarity configuration. The dimensionalities of these configurations were chosen according to the elbow criterion[2]. Each histogram represents the procrustes errors from $10,000$ comparisons of random configurations with the same dimensionalities as the empirical ones. The 8 comparisons in which the cutoff was not met, arose from 3 categories. The two-dimensional configuration for the item-item similarities of the ***tools*** category was responsible for 5 of these 8. This result might point towards a poor representation of the similarities in only two dimensions. (According to the elbow criterion the representation of these similarities requires three dimensions.)

Discussion

In the preceding sections we have put the assumption that categorization decisions for different items are made independently from one another to the test. The Rasch model, with its assumption of independence, was applied to categorization

[1] MDS yields an error measure called stress that expresses the deviance of the estimated distances from the observed associations. Empirically observed stress values can be compared against stress values that were obtained for random data to establish whether there is any structure present at all in the data that is being scaled. Structured data are expected to present with lower stress values than data without any real structure. This is of particular importance to the dependency estimates for which we took up the task of demonstrating that they contain information that has traditionally been overlooked. We subjected $10,000$ sets of uniformly distributed random associations between 24 items to the same nonmetric MDS procedure the dependency estimates were subjected to. The lowest stress value that was obtained in 2 dimensions equaled .294. In dimensionalities 3 to 6 the lowest stress values equaled .204, .140, .109, and .088, respectively. The stress values obtained for the dependency estimates were invariably lower than the lowest values obtained through scaling of the random data. The dependency estimates can thus be discerned from random data. If the independence assumption would be justified, the correlations among the residuals following the application of the Rasch model should not present with structure. Clearly, the above simulations support our earlier findings that they do. In addition, the simulations do not suffer from the multiple comparison problem, which is an issue in the significance testing of the correlations between residuals.

[2] When the stress values are plotted as a function of dimensionality, the resulting curve may exhibit a noticeable elbow. From that dimensionality onward the decrease in stress is thought to reflect error fitting. The elbow is therefore believed to indicate the “appropriate” dimensionality. See Verheyen et al. (2007) for details.

data for eight natural language categories. The subsequent investigation of the residuals learned that departures from the independence assumption indeed occur. Accordingly, these dependencies should be explicitly acknowledged in the various frameworks that are being used to explain categorization behavior. The Rasch model is part of the class of item response models. Tuerlinckx and De Boeck (2004) have proposed a range of item response models that allow LIDs to be taken into account. We will focus here on a restricted version of what they term a nonrecursive model for LIDs. In this model dependencies between items are handled by expressing the probability that a positive response is given to item *i* conditionally on the responses to all other items:

$$\text{logit}\left(p_{ci} \mid x_c^{(i)}\right) = \beta_i - \theta_c + \sum_{j \neq i} x_{cj}\delta_{ij} \qquad (4)$$

where the vector $x_c^{(i)}$ contains all responses of participant c except the one to item *i*. We assume parameter δ_{ij} to equal δ_{ji}.

Based upon the results reported earlier, we expect that item-item similarity would provide the most promising interpretation of the δ parameters in Equation (4). If this model were to be found appropriate for categorization data and if its parameters were indeed to correlate with similarity judgments, the model would be the first one that we know of that allows similarities to be computed from categorization data instead of the other way around. We believe this would constitute quite an achievement as the structural information carried by similarities is usually required by categorization models to properly predict categorization decisions. Indeed, most categorization models - like the Generalized Context Model (GCM, Nosofsky, 1984) for instance - require at their input the similarities of the items that are to be categorized. In the GCM this input takes the shape of a multidimensional space of which the dimensions are elongated or shrunken appropriately to allow correct categorization. Several well-motivated processes drive these transformations of the input space, but it is clear that in natural language categories much of the categorization burden is achieved through the use of the already highly structured similarity space. We do not wish to detract from the merits of the GCM and other similar categorization models, but at the very least it is a bit odd that these similarities are regarded as mere input, external to the model, and are therefore not taken into account when the complexity of the model is determined. Contrary to this, we hope to construct a model of categorization that is explicit about the parametric complexity it entails and does not require similarity as its input. To the contrary, if the model would prove able to predict item-item similarities, this would accommodate the long standing concern that traditional models of categorization are unjustified in using similarity as an explanatory construct (Goodman, 1972; Murphy & Medin, 1985) and would support authors like Quine (1977) who have suggested that it could be the knowledge that two items are or are not in the same category that drives judgments about their similarity.

References

Andrich, D., & Kreiner, S. (2010). Quantifying response dependence between two dichotomous items using the Rasch model. *Applied Psychological Measurement*, *34*, 181-192.

Borg, I., & Groenen, P. (2005). Procrustes procedures. In I. Borg & P. Groenen (Eds.), *Modern multidimensional scaling* (p. 429-448). New York, NY: Springer.

Goodman, N. (1972). Seven strictures on similarity. In N. Goodman (Ed.), *Problems and projects* (p. 437-447). Indianapolis, IN: Bobbs-Merrill.

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

Lunn, D. J., Thomas, A., Best, N., & Spiegelhalter, D. (2000). WinBUGS: A Bayesian modelling framework: Concepts, structure, and extensibility. *Statistics and Computing*, *10*, 325-337.

Murphy, G. L., & Medin, D. L. (1985). The role of theories in conceptual coherence. *Psychological Review*, *92*, 289-316.

Nosofsky, R. M. (1984). Choice, similarity, and the context theory of classification. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *10*, 104–114.

Quine, W. V. O. (1977). Natural kinds. In S. P. Schwartz (Ed.), *Naming, necessity, and natural kinds* (p. 155-175). Ithaca, NY: Cornell University Press.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen, Denmark: Danish Institute for Educational Research.

Rips, L. J. (1975). Inductive judgments about natural categories. *Journal of Verbal Learning & Verbal Behavior*, *14*, 665-681.

Rumelhart, D. E., & Abrahamsen, A. A. (1973). A model for analogical reasoning. *Cognitive Psychology*, *5*, 1-28.

Smith, J. D., & Minda, J. P. (1998). Prototypes in the mist: The early epochs of category learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *24*, 1411–1436.

Tuerlinckx, F., & De Boeck, P. (2004). Models for residual dependencies. In P. De Boeck & M. Wilson (Eds.), *Explanatory item response models: A generalized linear and nonlinear approach* (p. 289-316). New York, NY: Springer.

Verheyen, S., Ameel, E., & Storms, G. (2007). Determining the dimensionality in spatial representations of semantic concepts. *Behavior Research Methods*, *39*, 427-438.

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.

Wright, B. (1977). Solving measurement problems with the Rasch Model. *Journal of Educational Measurement*, *14*, 97-116.