Running head: THE PROBABILISTIC THRESHOLD MODEL

The Probabilistic Threshold Model

Steven Verheyen

University of Leuven, Leuven, Belgium

James A. Hampton

City University, London, England

Gert Storms

University of Leuven, Leuven, Belgium

Address for correspondence:

Steven Verheyen

Department of Psychology

University of Leuven

Tiensestraat 102, Bus 3721

BE-3000 Leuven

e-mail: steven.verheyen@psy.kuleuven.be

Abstract

A formalization of the Threshold Theory, called the Probabilistic Threshold Model, is introduced. According to the Threshold Theory semantic categorization decisions come about through the placement of a threshold criterion along a dimension that represents items' similarity to the category representation. The adequacy of this theory is assessed by applying the Probabilistic Threshold Model to categorization data for eight natural language categories and subjecting it to a formal test. In validating the model special care is given to its ability to account for inter- and intra-individual differences in categorization and their relationship with item typicality. Extensions of the Probabilistic Threshold Model that can be used to uncover the nature of category representations and the sources of categorization differences are discussed.

Keywords: Categories; Graded membership; Typicality; Similarity; Rasch model

PsycINFO classification: 2260 Research Methods & Experimental Design

The Probabilistic Threshold Model

Introduction

From the work of McCloskey and Glucksberg (1978) it is known that natural language categories do not have fixed extensions. That is to say, people disagree on the items they are willing to endorse as category members and individuals do not endorse the same items on different occasions. This is particularly true of items that are moderately typical of the category. To the question of whether a *parachute* is a member of the ***vehicle*** category, for instance, 52% yes and 48% no responses were given in the McCloskey and Glucksberg study. Thirthy percent of the respondents changed their answer from the first to the second categorization session. An independent group of judges awarded the item an average typicality score of 4.38 out of 10. For items at the extreme ends of the typicality scale, respondents categorized much more consistently. All respondents agreed that the highly typical *car* is a ***vehicle***, while no one made a similar claim for the atypical *apartment*. For these items categorization decisions did not change from one session to the other, either.

According to the Threshold Theory (Hampton, 1995, 2007) a categorization decision for a particular item comes about through the assessment of the similarity of the item's representation to the category's representation. If the assessed similarity exceeds a certain threshold, the item is endorsed as a category member; otherwise it is not. The Threshold Theory accounts for the variable extension of categories by assuming that the threshold criterion can vary from one person to the other or from one occasion to the other (Hampton, 1995).

The Threshold Theory also reconciles the apparently contradictory finding (Osherson & Smith, 1997) that items afford both a binary membership decision and a

continuous typicality judgment. The theory states that both phenomena arise from a single underlying dimension (i.e., similarity to the category representation) that is common to all respondents. The relationship between typicality and this similarity dimension is assumed linear. Every single respondent is assumed to make a binary cut along the dimension, separating category members from non-members. Since these cuts are all made somewhere along the same dimension, averaging across respondents' binary decisions (1 for *yes*, 0 for *no*) results in a continuous degree or probability of membership measure. Because of inter-individual differences in the placement of the threshold criterion along the dimension a monotonically increasing relationship (bounded between 0 and 1) between similarity and degree or probability of membership is thus assumed (Hampton, 2007).

Hampton (1998) provides support for the Threshold Theory's assumptions by reanalyzing the McCloskey and Glucksberg (1978) data. McCloskey and Glucksberg had one group of participants categorize 492 items in 18 categories. Another group was asked to provide typicality ratings for the same material. Averaging the binary membership decisions across respondents resulted in a probability of membership measure that displayed the hypothesized relation with the average typicality ratings (i.e., similarity to the category representation). Probability of category membership monotonically increased with typicality, starting at a probability of zero at the low end of the typicality range, demonstrating a profound rise among the moderately typical items, and attaining a probability of one at the high end of the typicality range. The resulting curves for individual categories were very similar in shape to the ones in Figure 2 (to be presented later) that were obtained in a study that we conducted ourselves and will be discussed in detail below. Furthermore, Hampton (1998) established the correlation between average typicality and normalized membership probability (a transformation of membership probability that would show a straight line function with typicality if the membership curve followed the cumulative normal distribution function) at .93 (across all categories).

The above procedure clearly supports the Threshold Theory, but is limited as a formalization thereof. For one, it does not incorporate all of the Threshold Theory's assumptions. The procedure is, for instance, agnostic as to whether all respondents' decisions actually display the structure the Threshold Theory proclaims. Is the probability of endorsing atypical items as category members lower than the probability of endorsing typical items in all individuals? Or is there a significant group of respondents for which the probability of endorsing an item as a category member is the same, regardless of typicality? Such divergences would not be picked up by the Hampton (1998) procedure if the remaining participants *were* to adhere to the Threshold Theory assumptions. These divergences would be lost in the averaging process instead. Related to this issue is perhaps the greatest shortcoming of the Hampton (1998) procedure: It lacks a clear counterpart for the threshold criterion notion. Individual respondents' categorization criteria are not made explicit, making it difficult to test hypotheses regarding the sources of individual differences therein. For instance, it is not clear how one would go about testing whether two groups of categorizers employ a different threshold along a common scale based on their respective membership probability curves.

In what follows we will note the commonalities between the Threshold Theory and the Rasch model (Rasch, 1960; Thissen & Steinberg, 1986). We will argue that this item response model incorporates many of the assumptions made by the Threshold Theory and hence allows for a rigorous formal test of them by applying the model to an extensive categorization data set. We will also determine whether the model is able to account for the semantic categorization phenomena that were discussed by McCloskey and Glucksberg (1978). More specifically we will verify whether the model can account for inter- and intra-individual differences in categorization and their relationship with item typicality. We will conclude by discussing the manners in which the model can be employed to test hypotheses regarding the nature of category representations and the sources of

categorization differences.

The Probabilistic Threshold Model

The Rasch model (Rasch, 1960; Thissen & Steinberg, 1986) is an item response model of which the properties are well understood. It was developed within the context of aptitude testing where it is employed to estimate individuals' proficiency with regard to a number of questions of varying difficulty. It models the probability that person p endorses item i. It does so by awarding both persons and items a position along a common, latent scale. Their relative position then determines the probability of endorsement. Person p's position along the scale is indicated by θ_p. In the context of semantic categorization θ_p can be understood to represent the person's threshold criterion or the degree of liberalness/conservatism the person displays when making categorization decisions. Each item i's position along the scale is indicated by β_i which in the current context would represent the item's similarity to the category representation. The difference between the two positions (i.e., the value of $\beta_i - \theta_p$) determines the probability that p will endorse i as a category member:

$$\Pr(Y_{pi} = 1) = \frac{e^{\alpha(\beta_i - \theta_p)}}{1 + e^{\alpha(\beta_i - \theta_p)}} \tag{1}$$

Equation (1) expresses that the more β_i exceeds θ_p on the latent scale, the higher the probability of endorsing the item is, and vice versa. This is reminiscent of the Threshold Theory's claim that categorization decisions come about through the assessment of the similarity of the item's representation to the category's representation. This assessment results in the positioning of the items along a latent similarity scale (i.e., fixing the items' β_i values). The further along the scale an item is positioned, the higher its similarity to the category representation is assumed to be.

According to the Threshold Theory a threshold criterion is then imposed on the

scale to determine whether the assessed similarity affords a positive rather than a negative categorization decision. We take the value of θ_p to indicate the position of this threshold criterion. The probability expressed in Equation (1) decreases with θ. Low values of θ_p indicate rather liberal categorizers for whom a modest degree of similarity suffices to conclude category membership. High values of θ_p characterize more conservative categorizers who require extensive similarity between item and category to conclude category membership. We thus take the differences in the estimates of θ_p to correspond to the inter-individual differences in the placement of the categorization threshold criterion the Threshold Theory proclaims.

In Equation (1) an individual's response is not said to be deterministic. Instead, the Rasch model expresses the *probability* with which an individual will endorse a particular item. Depending on the relative difference between the corresponding θ_p and β_i values this probability will differ. The resulting probability curve takes an S-shaped form, starting of at a zero when the $\beta_i - \theta_p$ difference is large and negative, demonstrating a profound increase for small difference between β_i and θ_p, and leveling off again when the difference grows large and positive[1]. In this respect the model deviates from the original Threshold Theory in which the threshold acts as a decision boundary that rigorously separates members from non-members. According to the original theory items are, without exception, classified as category members when their similarity to the category representation surpasses the person's threshold. Items whose similarity does not surpass this threshold are not endorsed as category members. In the modeling framework we propose the difference between the values of β_i and θ_p determines the *probability* that p will endorse i as a category member. To highlight this difference with the original Threshold Theory, we will term the model the Probabilistic Threshold Model.

The probabilistic nature of the model becomes clear in Figure 1 that displays the positions of a single person (θ_p) and two items (β_i and β_j) by means of tic marks along

the horizontal axis. The items clearly differ with respect to their similarity to the category representation. Item i is less similar to the category than item j which is evidenced by the former being located lower on the common scale than the latter. The threshold of person p lies in between the two items. The black curves in the figure indicate how the probability of endorsing the items as category members changes as a function of θ. As β_j surpasses θ_p item j has a high probability of being endorsed by p. The dotted vertical line at position θ_p in Figure 1 crosses the black response curve associated with β_j close to a categorization probability of 1. β_i does not surpass θ_p and therefore has a low categorization probability associated with it. The dotted vertical line at position θ_p crosses the black response curve associated with β_i close to a categorization probability of 0.

****

Figure 1 about here.

****

Now imagine a person whose θ is located to the left of both items. β_i and β_j then surpass the categorization threshold and both would have high categorization probabilities associated with them. Imagine a person whose threshold is located much further along the latent similarity scale than that of person p. In fact, the degree of similarity this person requires to favor a positive membership decision is that high that neither β_i nor β_j surpasses the corresponding θ. Inspection of Figure 1 confirms that both items would have a low probability of being endorsed.

The Probabilistic Threshold Model thus allows a categorization decision to be considered the outcome of a chance experiment of which θ_p and β_i are the parameters. If we assume these parameters to remain the same, multiple repetitions of the experiment will not always result in the same categorization decision. With each repetition of the experiment, the probability that a person with a particular θ_p value will endorse an item

with a particular β_i value as a category member is given by Equation (1). The converse probability represents the probability that the person will not endorse the item as a category member. Depending on the values of these probabilities, a particular categorization response might be more or less suspect to change from one occasion to the other. In the item response models literature this interpretation of the probabilities associated with θ_p is known as the stochastic subject interpretation. It opens up the possibility to have the Probabilistic Threshold Model account for intra-individual categorization differences without having to posit that they are due to changes in the persons' categorization threshold. Indeed, from McCloskey and Glucksberg (1978) it is known that the vagueness of semantic categorization may be seen in both intra- and inter-individual categorization differences. In their study, different participants did not agree on the items that could be considered category members, but individual respondents also changed their mind when they were queried about the same items a month later. Within the Threshold Theory framework differences in threshold location are thought responsible for both kinds of differences (e.g., Hampton, 1995). Like the Threshold Theory, the Probabilistic Threshold Model associates inter-individual categorization differences with differences in threshold criteria (i.e., θ_p values). The model provides a different account of intra-individual categorization differences, however. The theory is agnostic as to why participants would employ a different threshold criterion in a categorization session that is only different from the previous one in that it is organized one month later. In its current form it would have to rely on extraneous justifications to provide a satisfying account of intra-individual categorization differences. The Probabilistic Threshold Model, on the other hand, offers an inherent explanation of these categorization differences by positing that the process that underlies categorization decisions is probabilistic in nature. Although the conversion of the deterministic Threshold Theory into the Probabilistic Threshold Model might thus involve a deviation

from the original, we believe this to be warranted since it promises to address both inter- and intra-individual categorization differences in a single framework without harm to the original theory's interpretation or need to rely on extraneous justifications.

In the following section we will introduce a categorization study involving 8 natural language categories with 24 items each. Because the nature of the items that are to be categorized might vary across the different categories, a different scaling of the response functions (the black curves in Figure 1) might be required for each category. To this end we have included a parameter α in Equation (1) that is constant across all items of a category but can vary from one category to the other. The Probabilistic Threshold Model as expressed in Equation (1) will be fit to the data from the categorization study to assess its appropriateness. This will determine whether categorization decisions indeed come about through the placement of threshold criteria along a latent scale. To verify whether the interpretation of the latent scale in terms of items' similarity to the category representation is justified, the correlation of the resulting β_i's with typicality ratings provided by independent participants will be calculated. If typicality can be assumed to increase linearly with similarity and the latent dimension can be interpreted as a specific category's similarity scale, a category's β_i's should correlate strongly with that category's typicality ratings. The availability of typicality data also allows us to establish whether the Probabilistic Threshold Model can account for the McCloskey and Glucksberg (1978) finding that inter- and intra-individual differences in categorization are most prevalent among items of intermediate typicality. This would further validate the Probabilistic Threshold Model and the interpretation of its parameter estimates and associated probabilities.

A more general version of the model in which a separate α_i is estimated for each item will also be fitted to the categorization data. This model is known as the two-parameter logistic model or 2PLM as it comes with two parameters (β_i and α_i) for

every item (Birnbaum, 1968). This allows the shape of the probability response curves of different items to differ from one another. The 2PLM wouldn't require the slope of the probability curves of β_i and β_j in Figure 1 to be the same, for instance. The 2PLM and the Rasch model are often used next to one another in the item response literature. The main reason for including the 2PLM in the current analyses is to verify whether any important deviations from the categorization patterns suggested by the Probabilistic Threshold Model exist that need to be substantiated. Possible explanations of such deviations have been proposed by Hampton (1998, 2010) and include the familiarity of the items, the ambiguity of the items, the believe that they can technically be considered category members or not, or the belief that membership is dependent on whether one takes the category in a broad or in a narrow sense. Similar systematic deviations that allow for a substantive interpretation have been found in other applications of the 2PLM within the semantic literature (e.g., Verheyen & Storms, 2010).

Method

Participants

Two hundred and ninety first year psychology students at the University of Leuven participated for partial fulfillment of a course requirement. Two hundred and fifty of them completed a categorization task. The remaining forty students provided typicality ratings. All participants were fluent speakers of Dutch.

Materials

Categories and items were taken from Hampton, Dubois, and Yeh (2006) who constructed 8 categories with 24 items each to study contextual influences on categorization. The categories consisted of two animal categories (***fish*** and ***insects***), two artifact categories (***furniture*** and ***tools***), two borderline artifact-natural-kind categories

(***fruits*** and ***vegetables***), and two activity categories (***sciences*** and ***sports***). The corresponding category items included both clear members, clear non-members, and borderline cases. All materials were translated into Dutch.

Procedure

The data collection took place in a large class room where all participants were present at the same time. Each of them was handed an eight page questionnaire to fill out. The students participating in the categorization task 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*. They were also given the opportunity to indicate that a particular item was unknown to them. The categorization task took about 15 minutes to complete.

The students participating in the typicality rating task were to indicate on a 7-point rating scale how typical they found the 24 items printed on each page to be of the category displayed on top. It was explained to them that high responses on the scale were to indicate that an item was *very typical* of the category, while low responses were to indicate that it was *very atypical*. They too were given the opportunity to indicate that a particular item was unknown to them. The typicality rating task took on average 20 minutes to complete.

Results

Participants rarely indicated that an item was unknown to them. A number of participants did omit responses without specifying that the corresponding items were unknown to them. Across the categorization and typicality rating task less than 2% of data points were missing. Two tailed t tests indicated that the number of missing responses did not correlate significantly with the average of the categorization or typicality ratings in either of the categories (all $p > .05$). Figure 2 holds the averaged

results of both tasks. For every category, the probability of making a positive membership judgment was plotted against the respective items' average typicality. Average typicality appeared to be a good predictor of categorization probability. A few exceptions notwithstanding, the probability of making a positive categorization decision increased with typicality. While the atypical and very typical items afforded decisions that are quite stable across participants, the decisions for the items of intermediate typicality were more volatile. This resulted in items that are atypical of the category receiving categorization probabilities that are close to 0, items that are of intermediate typicality receiving categorization probabilities that span almost the entire probability range, and highly typical items receiving categorization probabilities that are close to 1.

****

Figure 2 about here.

****

Averaged results that are similar to the ones that are presented here, have been taken to support the Threshold Theory in the past (Hampton, 1998). The notion of a threshold, which is a characteristic of individual categorizers, is absent in analyses that are conducted at the aggregated level, however. To lend credibility to the Threshold Theory we therefore analyzed the categorization decisions using the Rasch model (or the Probabilistic Threshold Model as we call it in the context of the semantic categorization). If the model can be shown to fit the categorization data, this would substantiate Threshold Theory's claim that categorization decisions come about through the placement of individual decision criteria along a latent scale that also holds the items. If the items' positions along this latent scale can then be shown to correlate with rated typicality, the proclaimed relationship between categorization and typicality can be said to hold at the level of individual participants, not just at the aggregated level.

Validating the Probabilistic Threshold Model

The Probabilistic Threshold Model was fit to each category's categorization data using specialized software for item response analyses. The R package **ltm** employs Marginal Maximum Likelihood Estimation (MMLE) to obtain estimates of β_i, θ_p, and α. For each of the eight included categories, 24 β estimates (one for every item i), 250 θ estimates (one for every participant p), and one estimate of α were thus obtained. In fitting the 2PLM to each category's categorization data, 24 α estimates (one for every item i) were obtained in addition to 24 β estimates and 250 θ estimates. Details of the R procedures can be found in Rizopoulos (2006). The typicality ratings, which were provided by an independent group of participants, were introduced after the model estimation to give a substantive interpretation of the β_i estimates.

It is important to note that the MMLE approach that was taken to estimate both models is just one of many procedures to have been proposed for item response model estimation (for an overview see Baker & Kim, 2004). As a test on our conclusions the Probabilistic Threshold Model and 2PLM were also estimated under a Bayesian approach using WinBUGS (Kim & Bolt, 2007; Lunn, Thomas, Best, & Spiegelhalter, 2000). The conclusions of these analyses were the same as the ones drawn following the MMLE analyses. The reported conclusions thus do not hinge upon the employed estimation procedure.

****

Table 1 about here.

****

One can compare the relative fit of the models to the categorization data using either the BIC statistic, the AIC statistic, or the likelihood ratio test. The BIC relative goodness of fit statistics for the various categories can be found in Table 1. For ***fruits***,

vegetables, ***furniture***, ***sciences***, and ***sports*** the BIC indicated that the improvement in fit does not warrant the extra parameters the 2PLM incorporates. The Rasch BIC for these categories was lower than the 2PLM BIC. For ***fish***, ***insects***, and ***tools*** the BIC suggested these additional parameters are warranted. The 2PLM BIC for these categories was lower than the Rasch BIC. The AIC statistic and the likelihood ratio test, on the other hand, indicated the 2PLM to be the relatively better fitting model for all eight categories. The three test statistics thus do not yield a uniform answer to the question of which model should be preferred. Therefore, to assess whether the Rasch model or the 2PLM is the more suitable model for the categorization data, an omnibus test (described in Tuerlinckx & De Boeck, 2005) was performed. Unlike the BIC statistic, the AIC statistic, and the likelihood ratio test, the omnibus test constitutes an *absolute* measure of fit. We therefore consider it an appropriate arbitrator in choosing between the Rasch model and the 2PLM.

The omnibus test entails a comparison of the deviance, defined as -2 times the natural logarithm of the maximum likelihood, that was obtained after fitting either the Rasch model or the 2PLM to the categorization data, with 100 replicated deviance values. These were obtained by simulating 100 replicated data sets according to the models' estimated parameters and re-fitting the model to these data sets. The resulting deviances are then used to estimate the p-value of a goodness-of-fit test as follows:

$$\hat{p} = \frac{1}{100} \sum_{j=1}^{100} I(dev_j^{rep} > dev^{obs}), \tag{2}$$

where dev_j^{rep} refers to the deviance obtained from the jth replicated data set and $I(C)$ is the indicator function taking value 1 if condition C is true and 0 otherwise. If a model fits the data, the observed deviance should not differ too much from the simulated deviances, resulting in a $\hat{p}$-value that is close enough to .50. The second column of Table 2 holds the obtained $\hat{p}$-values for the Rasch model. The third column holds those obtained for the 2PLM. The Rasch $\hat{p}$-values did not deviate strongly from .50 indicating a sufficiently good

fit of the Rasch model to the semantic categorization data. The 2PLM $\hat{p}$-values, on the other hand, were close to 1 indicating that the model's estimated parameters might provide a good fit to the empirically obtained data, but do not allow generalization to data sets that might have been obtained as well. We therefore believe the 2PLM to overfit the data and the Rasch model to be the preferred model for the categorization data[2].

****

Table 2 about here.

****

Since the omnibus tests that were performed constitute tests of the absolute fit of the Rasch model to the categorization data, the $\hat{p}$-values that were obtained can also be taken to indicate that semantic categorization occurs through the placement of decision criteria along a common, latent dimension by individual respondents. Along this dimension the various potential category members are organized. This is in line with the assumptions of the Threshold Theory. To lend further support for the Rasch model as a proper formalization of the Threshold Theory, the latent dimension on which the model situates both persons' criteria and items is to represent similarity to the category representation. If we assume a linear relationship between typicality and similarity as did Hampton (1998, 2007), this proves to hold. The correlation between β_i and average rated typicality was established at .96 for ***fruits***, .97 for ***vegetables***, .94 for ***fish***, .97 for ***insects***, .97 for ***furniture***, .97 for ***tools***, .94 for ***sciences***, and .98 for ***sports***. These were all very close to the maximum correlations afforded by the reliability of the typicality ratings. The split-half correlations with Spearman-Brown correction were estimated at .99, .99, .98, .98, .99, .99, .96, and .99, respectively.

Accounting for the McCloskey and Glucksberg (1978) findings

The aim of the previous section was to establish whether the Threshold Theory holds at the level of individuals making categorization decisions. By demonstrating that the Rasch model is an appropriate model for semantic categorization, we provided evidence for the Threshold Theory's claim that individuals' categorization decisions come about through the placement of an individual criterion on a scale that is common to all categorizers. A strong linear relationship between the items' locations along the scale and their rated typicality provided evidence that the similarity of the items' representation towards the category's representation is at the basis of semantic categorization.

The Rasch model differs in one important respect from the original Threshold Theory. It is probabilistic, rather than deterministic: The further along the scale an item is located from an individual's criterion, the higher the probability that the individual will endorse the item, and vice versa. Because of this we termed the model the Probabilistic Threshold Model. Before, we already mentioned that we believe the probabilistic nature of the model to be an asset in that it promises to account for intra-individual differences, next to inter-individual differences in semantic categorization. McCloskey and Glucksberg (1978) were the first to demonstrate the vagueness of semantic categorization through these differences. In this section we reiterate their findings and assess whether the Probabilistic Threshold Model is able to bring them about for the data set under study.

When individuals require a different degree of similarity before endorsing an item as a category member, inter-individual differences in categorization arise. McCloskey and Glucksberg (1978) demonstrated that disagreement among categorizers is the highest for items of intermediate typicality. They calculated the proportion of nonmodal categorization responses at each level of the 10-point typicality scale they employed. At typicality level 4 the proportion was the highest with a value of .36. The proportion of nonmodal responses dropped off quickly towards both ends of the typicality scale. The

same holds true for the categorization data under study. The proportion of nonmodal categorization responses was established at .02, .08, .28, .37, .27, .08, and .02 at the seven points of the typicality scale we employed.

In the Probabilistic Threshold Model every categorizer-item-pair is associated with a value between 0 and 1 expressing the probability that the particular categorizer will endorse the item she is faced with. The extent to which this probability deviates from 1 or 0 for items that on average will or will not be endorsed as category members, respectively, constitutes a direct measure of the probability of providing a nonmodal response. If the average categorizer is likely to endorse item i as a category member and the probability that categorizer c endorses i is .83, for instance, there is a 17% probability that she will provide the nonmodal response *'not a member'*. In order to account for the McCloskey and Glucksberg (1978) data on inter-individual differences in categorization, the probability of providing a nonmodal response predicted by the Probabilistic Threshold Model should drop off from items of intermediate typicality to items that are at the extreme ends of the typicality scale.

****

Figure 3 about here.

****

For each level of the typicality scale that we employed in our study, Figure 3 expresses the probability of a nonmodal response as predicted by the Probabilistic Threshold Model. To allow comparison with the results reported by McCloskey and Glucksberg (1978) the predicted probability was calculated across all categorizers and categories (in the manner that was demonstrated before). Nonmodal responses appeared most likely for items that were judged to be of intermediate typicality. The highest probability was predicted at typicality level 4 of the 7-point typicality scale. The

probability of a nonmodal response was estimated at .36 for this typicality level. At typicality levels 3 and 5 nonmodal responses were still likely to occur, with estimates of .28 and .26, respectively. At the atypical end of the scale the probability of a nonmodal response was much lower with probabilities of .07 and .02 for typicality levels 2 and 1. A similar drop was noticeable at the highly typical end of the scale. The probability of a nonmodal response was estimated at .08 for typicality level 6 and at .02 for typicality level 7. The results are in accordance with our own empirical findings and those by McCloskey and Glucksberg, indicating that the Probabilistic Threshold Model correctly predicts the occurrence of nonmodal responses to be a function of typicality, with inter-individual differences occurring more often among items of intermediate typicality than among atypical or highly typical items.

A similar conclusion was reached by McCloskey and Glucksberg (1978) for intra-individual differences in categorization. These were found to occur most often for items of intermediate typicality as well. The proportion of within-categorizers inconsistencies was determined to be the highest at typicality level 4 of 10 with a value of .22. Hampton et al. (2006) also studied within-categorizers inconsistencies. They used the same stimuli we employ here, to establish the proportion of inconsistencies at the middle of the typicality scale around .18. At none of the other levels of the typicality scale was this proportion found to be higher. As was indicated before, in the Probabilistic Threshold Model the probability associated with an individual encountering a particular item to categorize has a natural interpretation in terms of intra-individual differences. We referred to this interpretation of the probabilities as the stochastic subject interpretation. If the probability with which categorizer c will endorse item i is .77, for instance, the probability that she will provide the opposite response is .23. If we determine the probability that categorizers will deviate from their most likely response for all items and determine the average of these probabilities for each level of the typicality scale, we can verify whether

the Probabilistic Threshold Model correctly predicts the McCloskey and Glucksberg finding that these deviations are most likely for items of intermediate typicality. Note that the manner in which the probabilities of within-categorizers inconsistencies are derived is different from the procedure to obtain the probability of nonmodal responses. To determine the probability of a person's nonmodal response for item i reference was made to the dominant categorization decision *across categorizers*. To determine the probability of a person's inconsistency, we will make reference to that *person's* dominant response to i. In the above example the latter probability was estimated to be .23. If the dominant response across categorizers for item i would be to deny it as a category member, the probability of a nonmodal response by categorizer c would be estimated at .77.

****

Figure 4 about here.

****

Figure 4 shows the probability of a categorization inconsistency predicted by the model, averaged across persons and categories, for each level of the typicality scale. The Probabilistic Threshold Model expressed inconsistencies to be most likely for items of intermediate typicality. The highest probability was predicted at typicality level 4 with a value of .26. The probabilities at typicality levels 3 and 5 were somewhat smaller with estimates of .23 and .21, respectively. The probability of a categorization inconsistency then quickly dropped off toward both ends of the typicality scale. Atypical items were associated with an average inconsistency probability of .07 at typicality level 2 and .02 at typicality level 1. Typical items were associated with a low average inconsistency probability as well. At typicality level 6 this probability was estimated at .07. At typicality level 7 it was estimated at .02. The results are in accordance with the McCloskey and Glucksberg (1978) finding and its replication by Hampton et al. (2006).

The Probabilistic Threshold Model correctly identifies the items of intermediate typicality to be those for which categorization inconsistency is most likely.

In addition, the model correctly indicated the probability of within-categorizers inconsistencies to be lower than the probability of between-categorizers differences at the intermediate level of typicality. While in Figure 4 the maximum probability was estimated at .26, it was estimated at .36 in Figure 3. In McCloskey and Glucksberg (1978) the maximum proportion of inconsistencies was estimated at 22%, while the maximum proportion of nonmodal responses was estimated at 36%.

This finding allows one to compare the Probabilistic Threshold Model, which includes a separate threshold for every participant, to a related model that assumes all participants to employ the same threshold criterion[3]. In all other respects the Probabilistic Threshold Model and this dummy model are the same. The latter model was also fitted to the categorization data under study. Unlike the Probabilistic Threshold Model, it predicted the probability of inconsistencies to be as high as the probability of nonmodal responding. This is clearly not in line with the McCloskey and Glucksberg (1978) findings, where intra-individual differences were found to be less prevalent than inter-individual differences.

Investigating threshold criterion stability

The results above support the Threshold Theory in general, and the Probabilistic Threshold Model in specific, as a framework for the study of what one could call "traditional" semantic categorization behavior. With the model offering estimates of individuals' threshold criteria θ_p it also becomes possible to study aspects of semantic categorization that have been rather neglected. One could, for instance, investigate to what extent the degree of liberalness/conservatism exhibited by a person in one category generalizes to another. To this end we correlated the participants' θ_p estimates for the

eight natural language categories they were presented with. One-tailed t-tests indicated that 22 of these $(8 \times 7)/2 = 28$ correlations were significant at the .05 level of significance. A one-sample t-test on the to Fisher z's transformed correlations indicated these correlations to come from a distribution with a mean greater than zero ($t(27) = 9.97, p < .001,$ one-tailed). Despite the fact that the correlations were of moderate magnitude (the maximum correlation, between ***furniture*** and ***tools***, only reached .34) these results point toward a considerable amount of stability in categorization behavior. In what follows we will discuss how the Probabilistic Threshold Model can be extended to uncover sources of variability and stability in semantic categorization.

General Discussion

Because of differences in range and discriminatory power, typicality and degree or probability of category membership have been said to tap into fundamentally different aspects of conceptual representations (Osherson & Smith, 1997). The fact that graded membership is bounded between 0 and 1, while typicality is thought of as being unbounded, is generally taken to support this argument. The Threshold Theory (Hampton, 1995, 2007) contests these claims. It states that both notions relate to a single underlying dimension: Typicality is understood to increase linearly with a metric of similarity, while degree of category membership is assumed to increase monotonically with this similarity metric. Hampton (1998) has provided evidence for these assumptions at the aggregated level. If one averages across respondents' discrete categorization decisions, a continuous measure of category membership arises that increases monotonically with typicality. At the level of individual participants, however, the Threshold Theory has to account for the discrete categorization decisions that are made. It does so by assuming that participants place a threshold criterion on the similarity metric that distinguishes category members from non-members. As the similarity metric is assumed common to all

participants, they are all suspected to adhere to it when making categorization decisions: An item that is low in similarity to the category representation should be less likely to be endorsed than an item that is higher in similarity by every single respondent. Of course participants can differ in the degree of similarity they require to endorse an item (i.e. the placement of the threshold criterion). In fact, these individual differences are required for the monotonically increasing relationship of the averaged category membership measure with similarity and typicality to come about. If all respondents employed the same threshold criterion, averaging across their categorization decisions would result in a discrete measure of category membership instead of a continuous one. McCloskey and Glucksberg (1978) already established that there are inter-individual differences in categorization, especially for items that are of intermediate typicality for the category.

We advanced an item response model to formally assess whether the categorization decisions made by individual respondents adhere to the Threshold Theory assumptions. The Rasch model (Rasch, 1960; Thissen & Steinberg, 1986) is the formal equivalent of the Threshold Theory in that it assumes that categorization decisions come about through the placement of a decision criterion along a latent scale that is common to all categorizers. The model awards both categorizers and items a position along this scale. The relative position of categorizer and item determines whether or not the item will be endorsed. The more the item's position exceeds the categorizer's position along the scale, the more likely it becomes that the item will be endorsed, and vice versa. The categorizers' positions can therefore be understood as their threshold criteria, while the items' positions can be thought to reflect their similarity to the category representation. We termed the Rasch model with this interpretation of its parameters the Probabilistic Threshold Model.

The Probabilistic Threshold Model was applied to categorization data for eight natural language categories. For each of these data sets the model analysis yielded an underlying dimension along which both items and persons could be located. The items'

positions were shown to correlate strongly with average typicality as rated by independent judges. This validated the Threshold Theory at the level of individual categorizers. In addition, we showed that the Probabilistic Threshold Model demonstrated the relationship between inter-individual differences in categorization and typicality that McCloskey and Glucksberg (1978) had established. They found that nonmodal categorization responses were most prevalent among items of moderate typicality. The same was true when the expected proportion of nonmodal responses was expressed as the modeled probability that categorizers would deviate from the average categorization decision according to the Probabilistic Threshold Model.

McCloskey and Glucksberg (1978) established that intra-individual differences in categorization follow a similar pattern. Items of intermediate typicality are most likely to receive different categorization decisions on various occasions. While the original Threshold Theory explicitly accounted for inter-individual categorization differences by assuming that different categorizers employ different threshold criteria, it has not been that explicit about these intra-individual differences in categorization. In order to account for them the Threshold Theory would have to propose that individuals' threshold criteria change from one occasion to the other. This follows from its assumption that categorization involves a deterministic decision process that always results in items surpassing the threshold being endorsed as category members, and items falling below the threshold always being considered non-members. Although this could certainly be a valid position to take, one could also make the possibility of a change in categorization decision inherent to the relative position of the item and the person's threshold. This is the approach taken by the Probabilistic Threshold Model. The closer to each other the positions of threshold and item are estimated to be, the higher the probability of a categorization change. This is the case because in the Probabilistic Threshold Model a categorization decision is considered the outcome of a chance experiment constituted by

parameters θ_p and β_i. In the event that both parameters are estimated to be the same, Equation (1) establishes the probability of p endorsing i at .5, indicating that the categorization decision could go either way. This is also apparent in Figure 1. When θ_p is estimated to coincide with either β_i or β_j, the corresponding response functions indicate the categorization probability to equal .5. Consequently, it would be considered highly likely that person p would provide different answers on two categorization occasions. When the expected proportion of within-categorizers inconsistencies was determined according to this uncertainty that the Probabilistic Threshold Model associates with each categorization decision, it was found that it was most prevalent for items of intermediate typicality. Its ability to demonstrate the McCloskey and Glucksberg findings on inter- and intra-individual categorization differences lends further credibility to the Probabilistic Threshold Model.

Explanatory item response models

Our main endeavor here has been to establish whether or not the Probabilistic Threshold Model is a suitable model for semantic categorization behavior. Because of this, the presented work is of an exploratory nature: Item response models were applied to empirical data, their fit was assessed, and attempts were made to relate the constituting parameters to an external empirical measure. As the Rasch model constitutes the first formal instantiation of the Threshold Theory, we deemed such an exploratory approach warranted in order to establish the model's appropriateness. It is, however, also possible to take an explanatory approach in which the external empirical measures are brought *into* the models (De Boeck & Wilson, 2004). Item and/or person characteristics are then incorporated in the item response models to test whether they can account for the variability found among the item and person parameters, respectively.

Up until now we have been fairly quiet about the nature of the dimension that

underlies the participants' semantic categorization decisions. We noted that it expresses the similarity of the items towards the category (i.e., typicality), but we did not elaborate on the nature of the representations involved (e.g., stored exemplars, an abstracted central tendency, ... - see Komatsu, 1992 for an overview). If one had specific hypotheses about the measures that determine an item's position along the latent scale one could test these by expressing the β_i's as a linear combination of these predictors. For instance, according to the generalized polymorphous concept model, the similarity between an item and a category can be expressed as a weighted combination of the number of characteristic features shared by item and category, and the number of features that are distinct to the item (Dry & Storms, 2010). Dry and Storms demonstrated how both common and distinctive feature information play a role in the prediction of items' typicality ratings (i.e., item-category-similarity). If one would want to learn whether and to what extent this finding generalizes to categorization, one could do so by employing the linear logistic test model (Fischer, 1995; Janssen, Schepers, & Peres, 2004). This model can be considered an explanatory version of the Rasch model in that it expresses the β_i's as a linear combination of item predictors, in this case feature commonality (FC) and feature distinctiveness (FD):

$$\beta_i = \gamma_0 + \gamma_{FC} FC_i + \gamma_{FD} FD_i + \varepsilon_i \text{ with } \varepsilon_i \sim N(0, \sigma_\varepsilon^2) \tag{3}$$

with γ_{FC} and γ_{FD} expressing the effects of feature commonality and feature distinctiveness, respectively, and γ_0 taking the role of intercept. When incorporated in the Probabilistic Threshold Model, (3) would allow for a test of the generalizability of the positive contribution of common features and the negative contribution of distinctive features to item-category-similarity, found by Dry and Storms. One could also imagine using this model to test whether different kinds of categories differ with respect to the relative contribution of common and distinctive features (e.g., natural kinds versus

artifacts or concrete versus abstract categories; see also Stukken, Verheyen, Dry, & Storms, 2009).

To further strengthen the proposed interpretation of β_i, we constructed an explanatory version of the Probabilistic Threshold Model in which β_i is regressed upon rated typicality in the same manner β_i is regressed upon feature commonality and feature distinctiveness in Equation (3). Where before the typicality ratings were introduced after the model had been estimated to aid interpretation of its parameters, the effect of rated typicality on the β_i's is now determined while the model is applied to the categorization data. In all eight categories typicality proved to be a significant predictor of the items' positions along the latent scale, accounting for 96% of the variance in β_i for the ***fruits*** category, 94% for ***vegetables***, 90% for ***fish***, 95% for ***insects***, 96% for ***furniture***, 95% for ***tools***, 90% for ***sciences***, and 97% for ***sports***. The estimated regression weight was positive in all categories, indicating that the higher an item's typicality, the further along the latent scale it would be found. These same conclusions had been reached earlier, based on the strong positive correlation between the (independent) β_i estimates and typicality ratings.

Inter-individual differences in categorization

Just as external variables can be brought into the Probabilistic Threshold Model to explain the variability among the items, so external variables can be brought into the model to elucidate the sources of variability among persons (Van den Noortgate & Paek, 2004). At the theoretical level there seems to be general agreement about semantic categorization behavior resulting from the interplay of the respondents' individual learning histories and the effects exerted by the immediate context they find themselves in (e.g., Barsalou, 1993; Smith & Samuelson, 1997). Empirical investigations into the sources of inter-individual differences in categorization are rare, however. Maybe this is the case

because up until now no principled way of determining individuals' degree of liberalness/conservatism in semantic categorization existed. The Probabilistic Threshold Model does offer such a measure in the form of the individuals' threshold criteria θ_p.

One could imagine employing the Probabilistic Threshold Model to investigate whether in semantic categorization there are systematic threshold shifts with age (Bjorklund, Thompson, & Ornstein, 1983; Lin, Schwanenflugel, & Wisenbaker, 1990). One only has to hear a child discuss her immediate environment to realize that the extensions of the categories she employs do not always match those held by adults. Overextensions are probably the most commonly found extension differences between children and adults. They occur when the child is excessively liberal in allowing items into a category (Clark, 1973). A child that refers to all four-legged animals as ***dogs***, for example, is making an overextension error. The reverse phenomenon - an excessively conservative use of a category label - is called an underextension (Nelson, 1974). It occurs when the child restricts the use of ***dog*** to German shepards only, for example. Underextension is much less likely to be noted than overextension in a child's spontaneous speech, since it involves the absence of a behavior. A categorization task like the one we employed in the current study might be ideally suited to detect underextension errors since it requires overt behavior from the child. If the Probabilistic Threshold Model could be fit to categorization data of various age groups simultaneously and systematic shifts (either more conservative or more liberal placement of the categorization threshold with age) in the person parameters θ_p could be shown to exist, considerable understanding of the manner in which children acquire categories can be achieved. Indeed, the compatibility of the Probabilistic Threshold Model and such developmental data would establish that young children already know about the single dimension that underlies categorization decisions, but do not yet agree with adults on the appropriate region to place the categorization threshold. If such results would be cast in the terminology of the previous

section, they could be taken to suggest that even young children know about the characteristic features that make up the underlying dimension, but do not accord them the same weight in categorization as adults do. This would corroborate existing theories of concept development (e.g., Johnson & Eilers, 1998; Mervis, 1984, 1987). Alternatively, shifts in item parameters across age groups would indicate a developmental reorganization of the category structure (see for example Keil & Batterman, 1984).

Similarly, one could imagine verifying whether a context manipulation induces a change in the employed threshold criterion. Braisby (Braisby, 1993, 2005) has noted that the clear context or purpose that helps shape natural discourse is generally absent in categorization tasks as they are performed in the psychology lab. He suggests that the variability in categorization decisions might arise from the lack of a clear context. Accordingly, Hampton et al. (2006) proposed a study in which categorization decisions were to be made in one of two clearly specified contexts. It was expected that in pragmatic contexts, people would take a broad view of what may be included in a category, whereas in technical contexts, the category boundary would be drawn more tightly. Contrary to these expectations, no difference was found between the conditions in the overall proportion of positive categorizations. In addition, categorization probabilities in each condition correlated equally strongly with typicality ratings that were provided by a group of participants who didn't receive a specified task context. It would be interesting to see whether an analysis using the Probabilistic Threshold Model might have a better chance of revealing context effects, as it takes the (typicality) structure of the data into account and does not carry the danger of obscuring possible important individual differences through aggregation.

Gardner (1953) was among the first to introduce the notion of a threshold in the context of categorization behavior and to study inter-individual differences therein. He argued for the existence of inter-individual criterion levels that remain stable across

different categorization tasks. In other words, a respondent who is found to be a rather conservative categorizer in one task, would also be expected to employ a rather strict criterion level in another categorization task. Following Gardner's suggestion we correlated the participants' θ_p estimates for the eight natural language categories they were presented with. The resulting correlations were found to come from a distribution with a mean greater than zero. This result lends support to Gardner's claim. Uncovering the personality characteristics that are responsible for the relative stability of categorization thresholds might therefore constitute another route to take the Probabilistic Threshold Model along in future research.

As a priori hypotheses concerning these relatively unexplored matters might be scarce, one might want to start by establishing whether there are groups of categorizers that differ substantially from one another. Rather than assuming that all categorizers are alike or that all categorizers are different from one another, one could look for latent groups of similarly performing categorizers. This is the approach taken by, among others, Lee and Webb (2005), Palmeri and Nosofsky (1995), Vanpaemel and Navarro (2007), and Verheyen and Storms (2007). It is straightforward to implement this demand for potential latent classes in the framework of the Probabilistic Threshold Model. Braeken and Tuerlinckx (2009) illustrate how to employ a finite mixture approach to solve the problem of determining the number of latent person groupings in estimating an item response model. One can imagine applying this procedure in the context of categorization to determine the number and nature of categorizer groups that are required. The model that is identified in this manner is situated somewhere between the Probabilistic Threshold Model, which includes a separate categorization criterion for every participant, and its dummy counterpart that assumes one criterion that is common to all participants. If the data permit, the Probabilistic Threshold Model may be employed in this manner to uncover groups of categorizers that adopt fundamentally different category

representations.

References

Baker, F., & Kim, S. H. (2004). *Item response theory.* New York, NY: Marcel Dekker.

Barsalou, L. W. (1993). Flexibility, structure, and linguistic vagary in concepts: Manifestations of a compositional system of perceptual symbols. In A. F. Collins, S. E. Gathercole, M. A. Conway, & P. E. Morris (Eds.), *Theories of memory* (p. 29-101). East Sussex, UK: Lawrence Erlbaum Associates.

Birnbaum, A. (1968). Some latent trait models. In F. M. Ford & M. R. Novick (Eds.), *Statistical theories of mental test scores* (p. 397-424). Reading, MA: Addison-Wesley.

Bjorklund, D. F., Thompson, B. E., & Ornstein, P. A. (1983). Developmental trends in children's typicality judgments. *Behavior Research Methods & Instrumentation*, *15*, 350-356.

Braeken, J., & Tuerlinckx, F. (2009). Investigating latent constructs with item response models: A MATLAB IRTm toolbox. *Behavior Research Methods*, *41*, 1127-1137.

Braisby, N. R. (1993). Stable concepts and context-sensitive classification. *Irish Journal of Psychology*, *14*, 426-441.

Braisby, N. R. (2005). Perspectives, compositionality, and complex concepts. In E. Machery, M. Werning, & G. Schurz (Eds.), *The compositionality of meaning and content (Vol. II: Applications to linguistics, psychology and neuroscience)* (p. 179-202). Frankfurt, DE: Ontos Verlag.

Clark, E. V. (1973). Whats in a word? On the child‘s acquisition of semantics in his first language. In T. E. Moore (Ed.), *Cognitive development and the acquisition of language* (p. 65-110). New York, NY: Academic Press.

De Boeck, P., & Wilson, M. (2004). *Explanatory item response models: A generalized linear and nonlinear approach.* New York, NY: Springer.

Dry, M. J., & Storms, G. (2010). Features of graded category structure: Generalizing the

family resemblance and polymorphous concept models. *Acta Psychologica*, *133*, 244-255.

Fischer, G. H. (1995). The linear logistic test model. In G. H. Fischer & I. W. Molenaar (Eds.), *Rasch models: Foundations, recent developments, and applications* (p. 131-155). New York, NY: Springer.

Gardner, R. W. (1953). Cognitive styles in categorizing behavior. *Journal of Personality*, *22*, 214-233.

Hampton, J. A. (1995). Testing the prototype theory of concepts. *Journal of Memory and Language*, *34*, 686-708.

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

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

Hampton, J. A. (2010). Stability in concepts and evaluating the truth of generic statements. In F. J. Pelletier (Ed.), *Kinds, things, and stuff: Concepts of generics and mass terms. New directions in cognitive science* (p. 80-99). Oxford: Oxford University Press.

Hampton, J. A., Dubois, D., & Yeh, W. (2006). Effects of classification context on categorization in natural categories. *Memory & Cognition*, *34*, 1431-1443.

Janssen, R., Schepers, J., & Peres, D. (2004). Models with item and item group predictors. In P. De Boeck & M. Wilson (Eds.), *Explanatory item response models: A generalized linear and nonlinear approach* (p. 189-212). New York, NY: Springer.

Johnson, K. E., & Eilers, A. T. (1998). Effects of knowledge and development on subordinate level categorization. *Cognitive Development*, *13*, 515-545.

Keil, F., & Batterman, N. (1984). A characteristic-to-defining shift in the development of word meaning. *Journal of Verbal Learning and Verbal Behavior*, *23*, 221–236.

Kim, J.-S., & Bolt, D. M. (2007). Estimating item response theory models using Markov Chain Monte Carlo methods. *Instructional Topics in Educational Measurement*, *26*, 38-51.

Komatsu, L. K. (1992). Recent views of conceptual structure. *Psychological Bulletin*, *112*, 500-526.

Lee, M. D., & Webb, M. R. (2005). Modeling individual differences in cognition. *Psychonomic Bulletin & Review*, *12*, 605-621.

Lin, P.-J., Schwanenflugel, P. J., & Wisenbaker, J. M. (1990). Category typicality, cultural familiarity, and the development of category knowledge. *Developmental Psychology*, *26*, 805-813.

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.

McCloskey, M. E., & Glucksberg, S. (1978). Natural categories: Well defined or fuzzy sets? *Memory & Cognition*, *6*, 462-472.

Mervis, C. B. (1984). Early lexical development: The contributions of mother and child. In C. Sophian (Ed.), *Origins of cognitve skills* (p. 339-370). Hillsdale, NY: Erlbaum.

Mervis, C. B. (1987). Child-basic object categories and early lexical development. In U. Neisser (Ed.), *Concepts and conceptual development: Ecological and intellectual factors in categorisation* (p. 201-233). Cambridge, UK: Cambridge University Press.

Nelson, K. (1974). Concept, word, and sentence: Interrelations in acquisition and development. *Psychological Review*, *81*, 267-285.

Osherson, D., & Smith, E. E. (1997). On typicality and vagueness. *Cognition*, *64*, 189-206.

Palmeri, T. J., & Nosofsky, R. M. (1995). Recognition memory for exceptions to the category rule. *Journal of Experimental Psychology: Learning, Memory, and*

Cognition, *21*, 548-568.

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

Rizopoulos, D. (2006). ltm: An R package for latent variable modeling and item response theory analyses. *Journal of Statistical Software*, *17*, 1-25.

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* (p. 161-195). East Sussex, UK: Psychology Press.

Stukken, L., Verheyen, S., Dry, M. J., & Storms, G. (2009). A new investigation of the nature of abstract categories. In N. A. Taatgen & H. van Rijn (Eds.), *Proceedings of the 31st Annual Conference of the Cognitive Science Society* (p. 2438-2443). Austin, TX: Cognitive Science Society.

Thissen, D., & Steinberg, L. (1986). A taxonomy of item response models. *Psychometrika*, *51*, 567-577.

Tuerlinckx, F., & De Boeck, P. (2005). Two interpretations of the discrimination parameter. *Psychometrika*, *70*, 629-650.

Van den Noortgate, W., & Paek, I. (2004). Person regression models. In P. De Boeck & M. Wilson (Eds.), *Explanatory item response models: A generalized linear and nonlinear approach* (p. 167-187). New York, NY: Springer.

Vanpaemel, W., & Navarro, D. J. (2007). Representational shifts during category learning. In D. McNamara & G. Trafton (Eds.), *Proceedings of the 29th Annual Conference of the Cognitive Science Society* (p. 1599-1604). Mahwah, NJ: Erlbaum.

Verheyen, S., & Storms, G. (2007). Modeling individual differences in learning hierarchically organised categories. *Psychologica Belgica*, *47*, 219-234.

Verheyen, S., & Storms, G. (2010). *Unidimensionality of category intension and extension.* Manuscript submitted for publication.

Author Note

Steven Verheyen is a research assistant at the Research Foundation - Flanders. This research was supported in part by grants G.0513.08 of the Research Foundation - Flanders and OT/05/27 of the Leuven University Research Council awarded to Gert Storms. We would like to thank Bieke Bollen, Simon De Deyne, Daniel Heussen, and Wolf Vanpaemel for their insightful comments on earlier versions of this manuscript. We are also in debt to Emmanuel Photos and two anonymous reviewers, whose valuable suggestions were very much appreciated.

Footnotes

[1]To attain this characteristic, item response models originally assumed a cumulative normal distribution. This was later changed to a logistic function so that model estimation would become easier. The Threshold Theory originally also assumed a cumulative normal distribution: Hampton (1998) employed a transformation of categorization probability that would show a straight line function with typicality if the membership curve follows the cumulative normal distribution (see our description earlier). This assumption is somewhat relaxed in Equation (1) that allows for a broader range of probability curves. The inclusion of the α parameter in Equation (1) allows to test whether this relaxation is required. If α is estimated to lie close to 1.702 the probability curves resemble a cumulative normal distribution. See in this respect footnote 2.

[2]The model in Equation (1) with α fixed at 1 has many applications in the item response literature. In the current context it can be set off against the Rasch model to verify whether it was worthwhile having a different α estimated for each category. All three relative goodness of fit statistics and the omnibus absolute goodness of fit test indicated that this was the case. The BIC, AIC, and likelihood ratio test indicated that the Rasch model with estimated α was the preferred model, except for ***fruits*** and ***vegetables***. A similar conclusion was reached based on the omnibus test. It indicated that the Rasch model with α fixed at 1 provided a sufficiently good fit to the categorization data for ***fruits*** and ***vegetables***, but not for the data of any of the other categories. Note that this does not impact on the conclusions reported in the text as α was of course estimated to lie close to 1 when the Rasch model in Equation (1) was fit to the ***fruits*** and ***vegetables*** categorization data. For none of the categories was α estimated to lie close to 1.702 (the maximum estimated α was 1.468). If it were, the logistic function relating the latent dimension to response probability would closely resemble a cumulative normal distribution.

[3]We would like to thank an anonymous reviewer for suggesting this comparison.

Table 1

BIC relative goodness of fit statistics for the Rasch model and the 2PLM.

	BIC	
Category	Rasch	2PLM
Fruits	3437.68	3500.12
Vegetables	3597.09	3641.92
Fish	4209.25	4071.85
Insects	4853.11	4838.51
Furniture	3720.40	3736.67
Tools	4221.28	4179.51
Sciences	4784.56	4809.06
Sports	3824.12	3842.51

Table 2

Omnibus absolute goodness of fit statistics for the Rasch model and the 2PLM.

	$\hat{p}$	
Category	Rasch	2PLM
Fruits	.62	.96
Vegetables	.70	.95
Fish	.58	.82
Insects	.69	.81
Furniture	.76	.96
Tools	.73	.92
Sciences	.46	.90
Sports	.53	.84

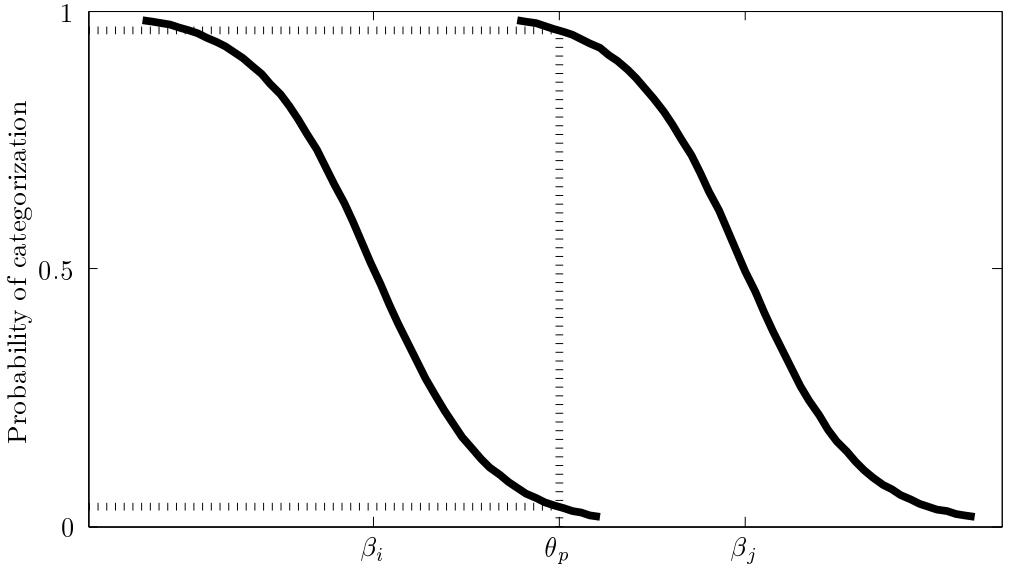
Figure Captions

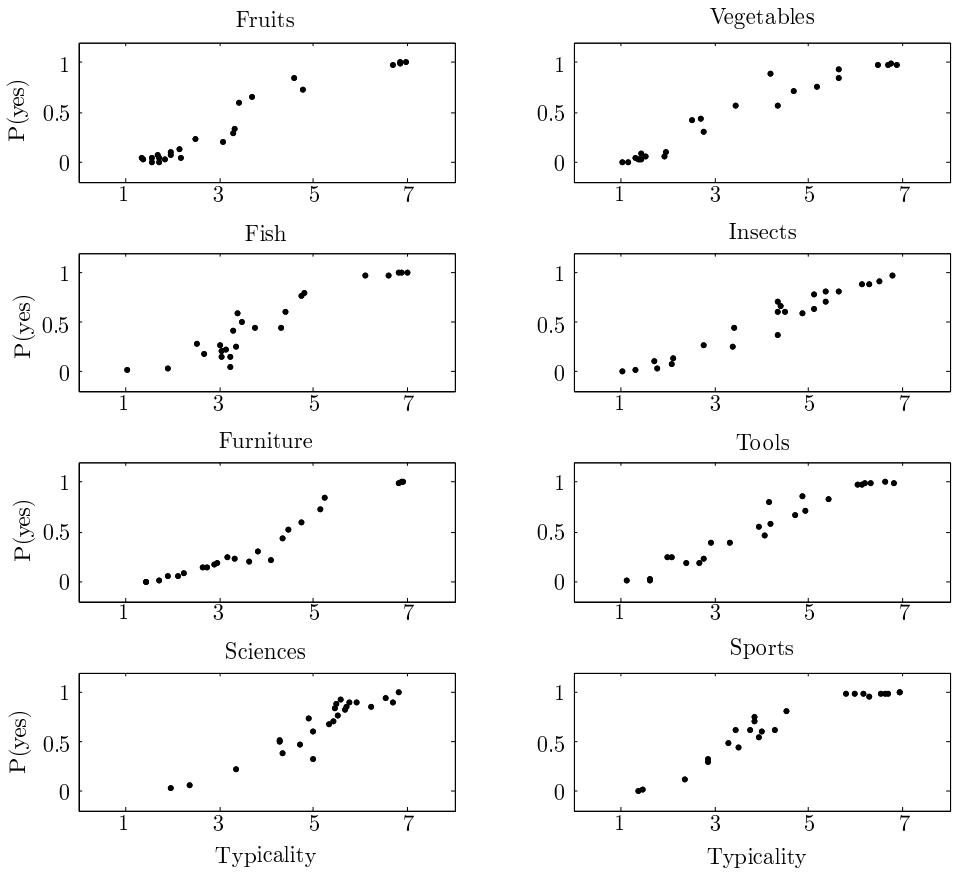
Figure 1. Illustration of the Probabilistic Threshold Model.

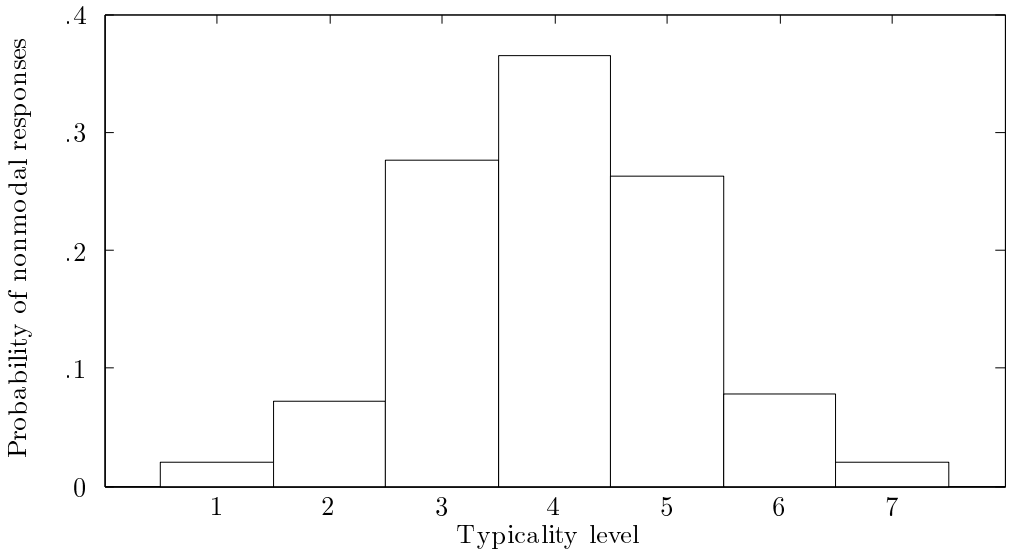
Figure 2. Scatterplots of the probability of a positive categorization versus average item typicality.

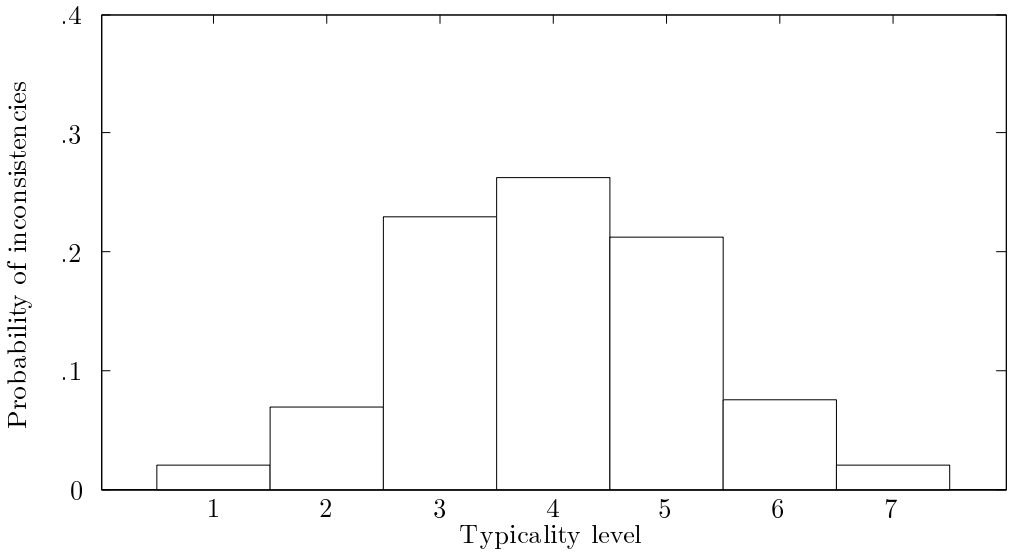
Figure 3. Probability of nonmodal responses as a function of typicality level.

Figure 4. Probability of within-categorizers inconsistencies as a function of typicality level.


1
0.5
0
Probability of categorization
β_i
θ_p
β_j


Fruits
Vegetables
Fish
Insects
Furniture
Tools
Sciences
Sports
P(yes)
Typicality
0
0.5
1
1
3
5
7


.4
.3
.2
.1
0
Probability of nonmodal responses
1
2
3
4
5
6
7
Typicality level


.4
.3
.2
.1
0
Probability of inconsistencies
1
2
3
4
5
6
7
Typicality level