Representation and Criterion Differences between Men and Women in Semantic Categorization

Loes Stukken (loes.stukken@ppw.kuleuven.be)
Steven Verheyen (steven.verheyen@ppw.kuleuven.be)
Gert Storms (gert.storms@ppw.kuleuven.be)
Faculty of Psychology and Educational Sciences. University of Leuven
Tiensestraat 102, 3000 Leuven, Belgium

Abstract

Gender differences are not widely studied in the categorization literature and the studies that did focus on gender differences generally investigated processing differences or differences in the use of particular categorization answers (absolute versus continuous). In the following study we looked at differences in the likelihood that men or women consider an item to be part of a category. The objective of the study was twofold: we wanted to introduce a model that is able to determine whether there are meaningful differences in categorization between groups and that is able to identify the sources of these differences. Secondly with this model we wanted to show that there were meaningful categorization differences between men and women: these differences are located at the level of the representation and/or the criterion.

Keywords: semantic categorization, threshold theory, gender differences, typicality, similarity, differential item functioning

Introduction

Are men more likely than women to consider **fishing** a *sport*? And are women more likely than men to consider a **dollhouse** a member of the category *toys*? Or in other words are there, for some items, differences between men and women in the likelihood that they would consider an item to belong to a particular category? And if so what is/are the source(s) of this gender difference? In the following study we addressed this question by gathering categorization judgments for 23 exemplars from eight categories, and by analyzing these data with a model that is able to detect differences between men and women in the strictness of the criterion they use to judge an item to be part of the category and differences in the representation that men and women use. The model is a random item mixture model proposed by Frederickx, Tuerlinckx, De Boeck, and Magis (2010), henceforth referred to as the RIM model.

RIM model

The RIM model is an item response theory model, Such models assume that the probability that a person endorses an item can be derived from the relative position of the item and the person towards each other on a common latent scale. The more an item's position exceeds the position of the person on the scale the more likely that the person will give a positive answer to the item. Verheyen, Hampton, and Storms (2010) claimed that these models therefore provide an excellent formalization of the threshold theory proposed by Hampton (1995, 2007) in which it is assumed that an item is judged to be part of a category if the similarity of the item to the category exceeds a certain threshold criterion. In this case the item's position on the latent scale represents the item's similarity to the category and the person's position is the threshold criterion the person uses to judge whether the item-category similarity is sufficient for the item to belong to the category.

The RIM model extends this approach in that it is able to account for group differences in two different ways. First of all, the RIM model estimates an average threshold criterion for each group. If the average threshold criterion estimated for women differs from the average threshold criterion for men, women have, depending on the sign of the difference, either a more liberal threshold criterion (they require a smaller item-category similarity than men to judge items as belonging to the category) or a more strict threshold criterion (they require a larger item-category similarity than men to judge an item to be part of the category). Thus the model allows us to detect whether the categorization differences between men and women are due to differences in the threshold criterion that they use to determine whether an item belongs to the category.

Secondly, the model is able to detect differential item functioning (DIF). An item demonstrates DIF when men and women who employ the same threshold criterion, nevertheless are found to have a different probability of endorsing an item. The RIM model allows the positions of these items on the latent scale to differ for different groups of people. The model is thus able to detect whether the position of an item on the latent scale should be different for men and women. The different position of the item indicates that the similarity of the item to the category differs between men and women and thus that men and women, given that they use the same threshold criterion, will have a different probability in judging this item to be part of the category. Different item positions for men and women thus imply representation differences between men and women. The RIM model is thus able to detect whether categorization differences between men and women are due to a difference in the representation and/or in the criterion between men and women.

The model is formally implemented by assuming that a categorization decision for item i by categorizer j from group g is the outcome of a Bernoulli trial with the probability that the item is judged to belong to the category equal to:

$$\text{logit}(\Pr(Y_{ijg} == 1)) = \beta_i - \theta_{jg}$$

In which β_i represents the item i's position on the latent scale and θ_{jg} represents the threshold criterion of categorizer j from group g. In a categorization context β_i can be taken to represent the similarity of item i to the category; θ_{jg} can be taken to represent the required level of item-category similarity to consider an item a category member (Verheyen, Hampton, & Storms, 2010). The model makes furthermore use of an indicator variable that indicates for each item, whether the item should be considered a DIF item. If so the model estimates a different β for that item for the two groups. If not the model estimates the same β for both groups.

Gender differences

Several studies showed that men and women differ in the processing of natural and artificial categories. Women tend to name and recognize members of natural categories faster, while men have an advantage over women in naming and recognizing artificial categories (Barbarotto, Laiacona, Macchi, & & Capitani, 2002; Capitani, Laiacona, & Barbarotto, 1999; Laws, 1999). Based on these studies Pasterski, Zwierzynska, and Estes (2011) argued that women and men might differ in the vagueness of their category judgments since natural and artificial categories tend to differ in vagueness. While membership in many natural categories is considered all-or-none, membership in most artifact categories is found to be graded (Diesendruck & Gelman, 1999; Estes, 2003, 2004; Verheyen, Heussen, & Storms, 2011).

Contrary to their initial hypotheses Paterski et al. showed that women provided more vague judgments than men (regardless of category type). They also showed that men, relative to women, gave more inclusive judgments for the artifact categories and tended to give more exclusive judgments for the natural categories.

Our study differs from these studies in that we are not focusing on differences between men and women in the processing of different types of categories or in the type of judgments that men or women give. We are interested in the question of whether or not there are differences in the likelihood/probability that men and woman judge an item to be part of the category. We argue that since men and women are known to be raised differently, to dress differently, to play with different toys, and to engage in different hobbies and professions, we expect that for some items men and women might differ in the likelihood that they consider the item as part of a particular category. To our knowledge this is the first study that looks at gender differences in the likelihood/probability that individual items are part of a category and allows to determine whether these differences reside at the level of the criterion or at the level of the representation.

Method

Materials

Eight natural language categories were studied (*Addictions, Clothing, Diseases, Furniture, Professions, Sports, Toys, Weapons*). The categories were selected based on the intuition of the researchers that they might contain items that have a different likelihood of membership in men and women. For each category we included 23 exemplars in the study. The items were selected based on previously collected typicality ratings to make sure that each category contained candidate exemplars that were generally considered typical of the category, atypical of the category, and borderline (items for which people in general are not always sure of whether they belong to the category or not). The typicality ratings were gathered as part of a larger norming project comprising 1276 items from 24 categories. Twenty-nine students (23 women, 6 men) provided typicality ratings for half of the categories using a seven point Likert scale ranging from *very atypical* to *very typical*. The reliability of these ratings for the 23 x 8 items in our study varied between 0.86 for *addictions* and 0.96 for *clothing* with a mean of 0.93.

Categorization task

In total 287 men and 568 women participated in the study. They filled in a questionnaire in which they were, for each item, asked to indicate (yes or no) whether it belonged to the corresponding target category. Participants were, for example, asked whether or not a **cold** was part of the category *diseases*. To prevent order effects, we administered 4 different versions of the task with a different order for items and categories. The participants were randomly assigned to one of the 4 versions of the task. The age of the participants ranged between 17 and 64 with an average of 20.

Model analyses

Each category's categorization data were analyzed separately using the RIM model. This was done using WinBUGS (Lunn, Thomas, Best, & Spiegelhalter, 2000) according to the details and code provided by Frederickx et al. (2010). For every analysis 5 chains were run of 10,000 iterations each, with a burn-in sample of 800.

Results

Typicality

For every category we calculated the correlation between the items' positions on the scale (the posterior means for the β_i's) and the items' average typicality to verify whether people were categorizing items by the use of similarity. We calculated the correlation between the items' positions and typicality because it was previously suggested that typicality and item-category similarity are strongly linearly related

(Hampton, 2007; Verheyen, Hampton, & Storms, 2010). A high correlation between typicality and the items' positions thus also implies a high correlation between the items' positions and the similarities of the items towards the category. We calculated the correlations between typicality and the items' positions for men and women separately. The correlations can be found in Table 1. The correlations were invariably high, suggesting that our participants were indeed using item-category similarity when they judged whether the items belonged to the category.

Table 1
For each category the correlation between item positions and typicality for men and women separately

Category	Men	Women
Addictions	0.92	0.92
Clothing	0.98	0.98
Diseases	0.92	0.95
Furniture	0.95	0.95
Professions	0.91	0.93
Sports	0.87	0.96
Toys	0.93	0.93
Weapons	0.94	0.94

Also note that the correlations of typicality with the items' positions of men and women can hardly be distinguished. Looking at categorization tendencies across the entire typicality range might not be the most fruitful manner to identify differences between groups of categorizers. For natural language categories, whose meaning is to a considerable extent determined by the environment the language community shares, one does not expect pronounced reorganizations of the representation from one group to the other. This would seriously hamper the communication between the group members. Rather, the differences might be more subtle, residing in individual items or in the severity of the employed categorization criterion.

Criterion differences

To check whether there were any gender differences in the threshold criterion that participants used to make category judgments, we plotted the posterior distribution of the difference in the average threshold criterion between men and women. If there is a reliable difference in the average threshold criteria, the credibility intervals of this distribution (the region around the mean that contains 95% of the mass of the distribution) may not include 0. As can be seen from Figure 1, this is the case for two categories: *professions* and *toys*. In these categories the average differences were 0.34 and 0.74 respectively, indicating that women had a more liberal threshold criterion and require less item-category similarity to judge items to be part of the category than men. For the other categories there is no credible difference in average threshold criterion indicating that women and men on average require equal levels of item-category similarity for category membership.

Representation differences

The RIM model gives an indication of the DIF-status of items by means of latent indicator values that can take one of two values (either DIF or no DIF) on every iteration, resulting in a difference in the estimated item position when required. Following Frederickx et al. (2010) we term an item a DIF item if in more than half of the iterations it was classified as DIF. Table 2 gives an overview of the number of items that were identified as DIF items and the number of items for which men seemed to be more inclined/likely to consider the item to be part of the category and the number of items for which women seemed to be more likely to judge the item to be part of the category. There was one category for which no DIF items were found, the category *furniture*. For one category, the category *clothing*, we found only one DIF item: **belt** was categorized differently by men and women with the same threshold criterion (men were more likely than women to indicate that **belt** was part of the category). For the other categories the number of DIF items ranged between 2 and 16 and for most of these categories there were both DIF items for which men were more likely to indicate that they were part of the category and DIF items for which women were more likely to indicate that they were part of the category. The categories *professions* and *weapons* were the only categories that contained only DIF items for which women were more likely to indicate that they were part of the category. For *professions* these items were **diver**, **magician**, **explorer**, **parachutist**, **pirate**, and **inventor**. For *weapons* these were **catapult** and **harpoon**.

DIF items were found across the entire range of typicality. Within the DIF items there were items for which people generally agree that it belongs to the category (for example: **dollhouse** for the category *toys***)**, items for which it is not sure whether or not they belong to the category (**snooker** for the category *sports*) and items for which it is generally agreed that they do not belong to the category **(pirate** for the category *professions*). Thus women and men do not only disagree on items for which there is uncertainty about whether or not they belong to the category, but also on items for which there is general agreement about whether or not they belong to the category.

First of all remember that for the category *clothing* the RIM model indicated that there is no reliable difference in mean threshold criterion between men and women. So any gender differences in categorization proportions are representation differences according to the model. The model considers only one of these differences meaningful. The model detected only one DIF item (**belt**, with an average typicality of 4.53**).**

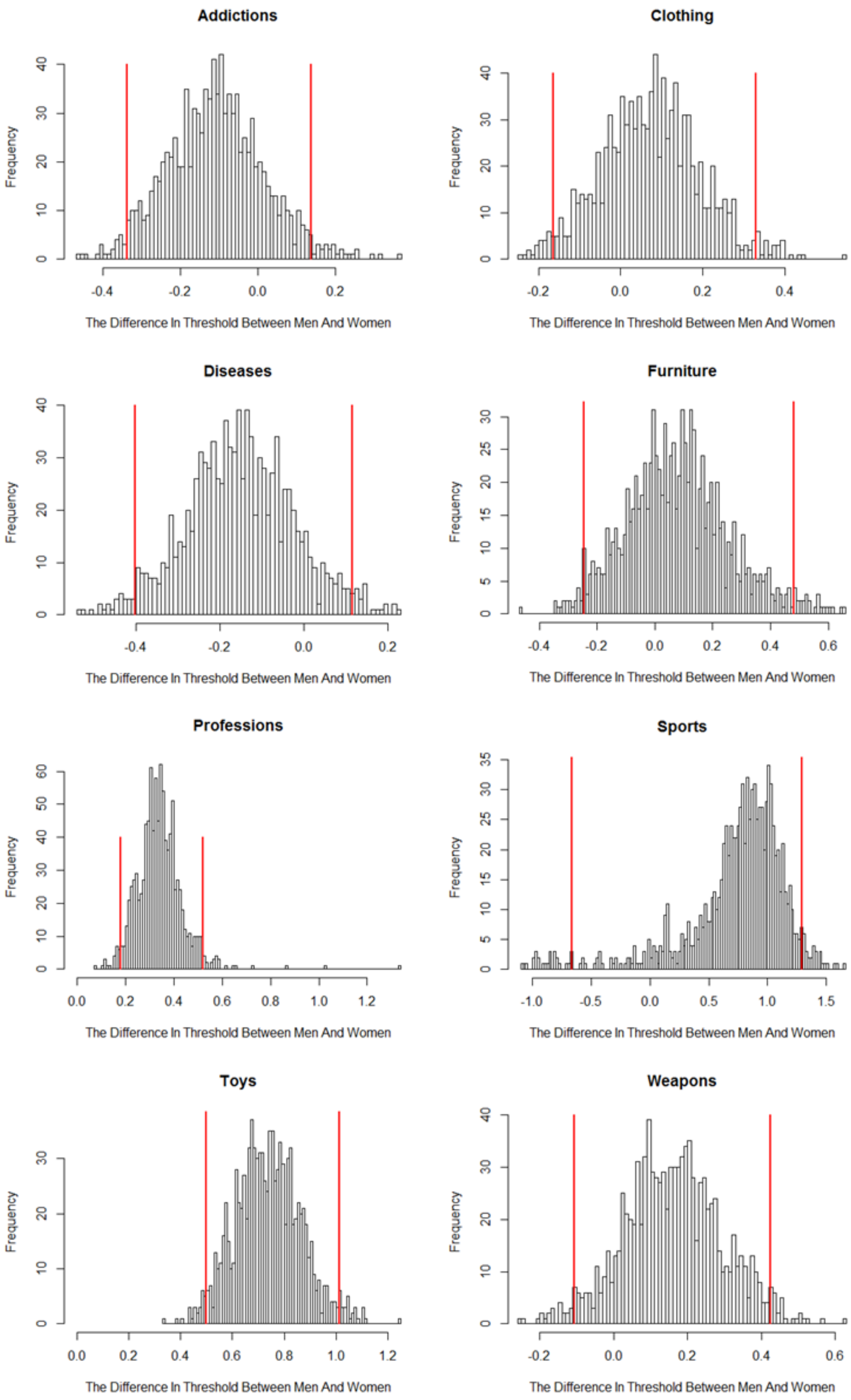

Figure 1: The posterior distributions of the difference in mean threshold between men and women for the eight categories. The 95% credibility interval is represented by the red bars.

Table 2
Overview of the number of DIF items in the categories

Category	# DIF items	Men[1]	Women[1]
Addictions	4	1	3
Diseases	10	8	2
Clothing	1	1	0
Furniture	0	0	0
Professions	6	0	6
Sports	16	5	11
Toys	6	4	2
Weapons	2	0	2

[1] columns 'Men' and 'Women' represent the number of DIF items for which respectively men and women are more inclined to consider it as part of the category

For the category *toys* there were several items for which the difference in proportion was determined meaningful after controlling for the threshold criterion. The RIM model indicated that the items **pin-ball machine**, **gocart**, **coloured pencil**, and **chalk** (manly items); and **dollhouse** and **skipping rope** (womanly items) are DIF items. That is, according to the model, the categorization differences one observes for these items are representational in nature. Interesting here are the items **comic book** and **music box**, that at first glance have a large and meaningful difference in the categorization proportion between men and women (0.60 versus 0.73 and 0.46 versus 0.62 at average typicalities of 4.23 and 5.08, respectively). The RIM model nevertheless indicates that these are not DIF items. After controlling for the threshold criterion there no longer is a meaningful difference between men and women for these items, indicating that the difference in proportion for these items is entirely caused by the difference in threshold criterion for this category. Indeed, the category of *toys* was one of the categories for which the RIM model indicated there was a credible criterion difference between men and women.

It is therefore also able to identify items for which at first glance there are no differences when one looks only at the differences in proportions between the groups/sexes. The item, **go-cart** (average typicality: 5.15), for example, has a very small difference in categorization proportion between men and women (0.78 versus 0.76), but after controlling for the threshold criterion the item is identified by the model as a DIF item.

These examples should make it clear that it does not suffice to look at categorization proportions alone to determine whether there are group/gender differences in categorization, and that the main contribution of the model is that it is able to disentangle two main causes of differences in categorization proportions: criterion differences and representation differences.

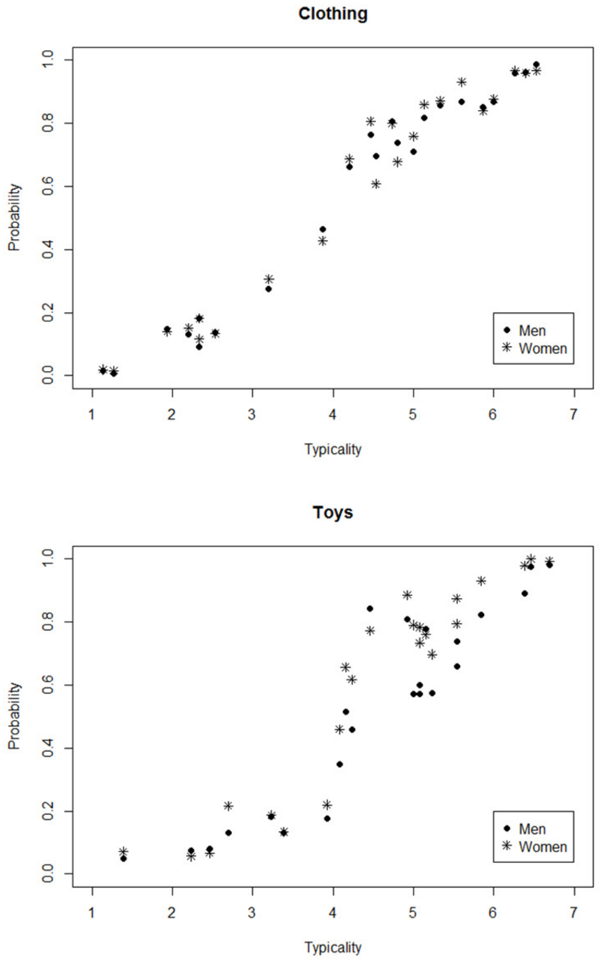

Figure 2: Categorization proportion for men and women as a function of typicality for the categories *clothing* and *toys*

Discussion

This is, to our knowledge, the first study that looks at differences in the likelihood/probability with which men and women would judge items to be part of the category. We have shown that for some items categorization differences were due to representational differences between men and women and for other items these differences were due to the differences in threshold criterion that men and women use.

The study described above fits in a recent group of studies in which it is shown that item response theory models can be used to analyze data from categorization tasks (Verheyen, Hampton, & Storms, 2010; Verheyen, De

Deyne, Dry, & Storms, 2011). It has previously been shown that variations of these models can reveal differences in the threshold criterion between different groups (Verheyen, Ameel, & Storms, 2011) and reveal latent groups of categorizers who employ a different representation (Verheyen & Storms, 2013). The RIM model is another valuable approach in that it can not only reveal criterion differences between existing groups, but also representation differences between these groups. It therefore opens up new ways of investigating group differences in semantic categorization.

For instance, a study by Hampton, Dubois, & Yeh (2006) that investigated categorization differences between groups of categorizers who were categorizing items in different contexts, compared the correlation between categorization proportions and typicality, and the percentage positive responses between the different groups. In our study there were only minor differences in the correlation between the categorization proportions and typicality between men and women, but the model did indicate that there were differences between the two sexes. Furthermore, looking at the percentage positive responses to see whether some groups are using a stricter threshold criterion, might give an imprecise picture of what is going on, since we showed that not all differences in categorization proportions are due to the use of a more or less strict categorization criterion. It should be clear from our results that the RIM model allows for the detection of more subtle but important differences between groups. The implementation of the model is also not limited to two groups. It can easily be extended to investigate data from multiple groups, such as the pragmatic, technical, and neutral context groups in Hampton, Dubois, & Yeh (2006), cultures (Medin & Atran, 2004), or age groups (Ameel, Malt, & Storms, 2008). The RIM model can also be easily adjusted to account for continuous categorization data.

Acknowledgments

Steven Verheyen is a postdoctoral fellow at the Research Foundation-Flanders. The authors wish to thank Sofie Frederickx for generously providing details about the RIM model.

References

Ameel, E., Malt, B. & Storms, G. (2008). Object naming and later lexical development: From baby bottle to beerbottle. *Journal of Memory and Language, 58*, 262-285.

Barbarotto, R., Laiacona, M., Macchi, V., & Capitani, E. (2002). Picture reality decision, semantic categories and gender: A new set of pictures, with norms and an experimental study. *Neuropsychologia, 40*, 1637-1653.

Capitani, E., Laiacona, M., & Barbarotto, R. (1999). Gender affects word retrieval of certain categories in semantic fluency tasks. *Cortex, 35*, 273-278.

Diesendruck, G., & Gelman, S. A. (1999). Domain differences in absolute judgments of category membership: Evidence for an essentialist account of categorization. *Pschonomic Bulletin & Review, 6*, 338-346.

Estes, Z. (2003). Domain differences in the structure of artifacual and natural categories. *Memory & Cognition, 31*, 199-214.

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

Frederickx, S., Tuerlinckx, F., De Boeck, P., & Magis, D. (2010). RIM: A random item mixture model to detect differential item functioning. *Journal of Educational Measurement, 47*, 432-457.

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

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

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

Laws, K. R. (1999). Gender affects naming latencies for living and nonliving things: Implications for familiarity. *Cortex, 35*, 729-733.

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

Medin, D.L & Atran, S. (2004). The Native Mind: Biological Categorization, Reasoning and Decision Making in Development Across Cultures. Psychological Review, 111(4), 960-983 .

Pasterski, V., Zwierzynska, K., & Estes, Z. (2011). Sex differences in semantic categorization. *Archives of Sexual Behavior, 40*, 1183-1187.

Verheyen, S., Ameel, E., & Storms, G. (2011). Overextensions that extend into adolescence: Insights from a threshold model of categorization. In L. Carlson, C. Hölscher, & T. F. Shipley (Eds.), *Proceedings of the 33rd Annual Conference of the Cognitive Science Society* (pp. 2000-2005) Austin, TX: Cognitive Science Society.

Verheyen, S., De Deyne, S., Dry, M. J., & Storms, G. (2011). Uncovering contrast categories in categorization with a probabilistic threshold model. *Journal of Experimental Psycholoy: Learning, Memory, and Cognition, 37*, 1515-1531.

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

Verheyen, S., Heussen, D., & Storms, G. (2011). On domain differences in categorization and context variety. *Memory & Cognition, 39*, 1290-1300.

Verheyen, S., & Storms, G. (2013). A mixture item response theory approach to vagueness and ambiguity. *Manuscript submitted for publication.*