Relationship between Extensions and Intensions in Categorization:

A Match Made in Heaven?

Farah Mutiasari Djalal[1]

James A. Hampton[2]

Gert Storms[1]

Tom Heyman[1]

[1] University of Leuven

[2] City, University of London

Correspondence address:

Farah Mutiasari Djalal

Faculty of Psychology and Educational Sciences

University of Leuven

Tiensestraat 102 Bus 3711

B-3000 Leuven

Belgium

Tel. 32 16 37 32 12

E-mail: farahmutiasari.djalal@kuleuven.be

Abstract

The present study investigated the relationship between category extension and intension for eleven different semantic categories. It is often tacitly assumed that there is a (strong) extension-intension link. However, a recent study by Hampton and Passanisi (2016) examining the patterns of stable individual differences in concepts across participants called this hypothesis into question. To conceptually replicate their findings, two studies were conducted. We employed a category judgment task to measure category extensions, whereas a property generation (in Study 1) and property judgment task (Study 2) were used to measure intensions. Using their method, that is, correlating extension and intension similarity matrices, we found non-significant correlations in both studies, supporting their conclusion that similarity between individuals for extensional judgments does not map onto similarity between individuals for intensional judgments. However, multi-level logistic regression analyses showed that the properties a person generated (Study 1) or endorsed (Study 2) better predicted her own category judgments compared to other people's category judgments. This result provides evidence in favor of a link between extension and intension at the subject level. The conflicting findings, resulting from two different approaches, and their theoretical repercussions are discussed.

Introduction

In studies of natural concepts, many theories have claimed that in order to categorize, people may recognize objects as having shared properties with other objects, and as a result group these objects together into the same category. For instance, if you encounter a novel object that has fur, four legs, a tail, and barks, you might compare it with objects that you know have similar properties, and then you could group this novel object together with other similar objects, in this case into the category of dogs. This example shows that concepts have two important aspects that play a role in categorization: the intension (the properties that define concepts) and the extension (the set of category members)[1].

It has been a long-held belief that category extension and intension are somehow related (e.g., Aristotle, 4th century BC/1961; Frege, 1948; for a recent overview see Hampton & Passanisi, 2016). As proposed by property-based models (Hampton, 1979; Rosch, 1975; Rosch & Mervis, 1975; Smith, Shoben, & Rips, 1974), the world is divided into natural categories that are structured by clusters of properties. People use these properties to make predictions and to draw inferences in deciding whether an object is a member of a category. Thus, an object will be grouped into a certain category if it has necessary and sufficient properties (i.e., the classical view) or shares certain properties with other members in the category (i.e., the probabilistic view, Smith & Medin, 1981). Such property-based models have been used in numerous studies of categorization, many of which conclude or assume that category extension and intension are closely related (e.g., Ameel, Malt, & Storms, 2014; Caplan & Barr, 1989; Hampton, Dubois, & Yeh, 2006; Murphy, 2002; Verheyen, De Deyne, Dry, & Storms, 2011).

[1] In philosophical semantics, the terms extension and intension refer to different sides of the same coin (i.e., they map to each other by definition). Here we follow the psychological literature in which extension usually refers to the list of category members that people endorse, whereas intension refers to the properties that they believe define concepts.

Moving in the opposite direction: Intension is not everything

Besides properties as the intensional information, other researchers focused on similarity to stored exemplars of categories in predicting category membership (Heit & Barsalou, 1996; Nosofsky, 1984). Storms, De Boeck, and Ruts (2000), for instance, found an exemplar-based model to be a better predictor of category membership than a property-based model. Furthermore, other studies have found that the relationship between extensions and intensions is not as direct as some have claimed. Malt, Sloman, Gennari, Shi, and Wang (1999) studied the names of containers in three different languages and argued that categorization can also be influenced by the linguistic and cultural histories of the language itself. More specifically, they proposed three mechanisms (i.e., chaining, convention, and pre-emption) that explain why properties alone may be insufficient for capturing the complexity of naming choices. These mechanisms may explain why object clustering is not solely based on a particular set of properties (see also Ameel, Malt, & Storms, 2008).

There are also accounts that presume a reverse relationship between exemplars and properties (Spalding & Gagné, 2013). In this view, category membership is a given rather than guided by the properties an exemplar possesses. Common or defining properties of a concept are *derived from* the exemplars of the category[2]. Despite their differences, all of these theories would expect some kind of (causal) relationship between category extension and intension.

However, Hampton and Passanisi (2016, henceforth H&P) recently called this assumption into question, suggesting instead that in one important respect intension might not map onto extension at all. They reasoned that *if* there is an intension-extension link (dis)similarity between individuals in terms of their category intensions should translate to (dis)similarity in their category extensions. Using a property importance rating task to

[2] We thank Thomas Spalding for bringing this to our attention.

measure category intensions and a typicality judgment task for category extensions, H&P examined whether individual variation in the representation of category extensions indeed maps to inter-individual variability in category intensions. That is, if two persons show similar typicality judgments, H&P argued that their judgments of property importance should also be similar. Conversely, if two persons weight properties differently, the idea is that their category extension (i.e., typicality ratings) should also diverge to a certain extent. For instance, if person A considers *physical activity* to be an important property of *sport*, whereas person B considers *having rules* to be more important, one would expect their typicality judgments to vary correspondingly. That is, *hiking* would be a more typical sport for person A, whereas *snooker* would be a more typical sport for person B. Analogously, if person A and C both consider *physical activity* to be an important property of *sport*, one would expect similar typicality ratings for *hiking* (i.e., relatively high ratings) and *snooker* (relatively low ratings).

To formally test this prediction, they constructed two similarity matrices for participants, one for extensions (i.e., consisting of all pairwise correlations between participants' typicality judgments) and a second one for intensions (i.e., consisting of all pairwise correlations between the same participants' property importance judgments). They then correlated these similarity matrices from the two tasks to test whether there is a relation between extensional and intensional representations. Across four studies (with slight variations in the methodology), they came to the conclusion that similarity between individuals in extensional judgments did not map onto similarity between individuals in intensional judgments. Estimates of the correlation between similarity matrices were close to zero in spite of test-retest reliability correlations for both matrices obtained in two of the studies of around .35, significantly above zero. Put differently, their results contradict the widely accepted view that intensions map to extensions. H&P proposed instead that category intension and extension are not integrated, but rather that they are stored independently from each other. More concretely, there might be an exemplar-based system underlying extensional judgments (see e.g., Storms, De Boeck, & Ruts, 2000) and a theory/schema-driven system underlying intensional judgments. Such a hybrid form of concept representation can account for H&P's findings by assuming that the typicality and property importance judgments tap into different systems, which are stored separately in semantic memory.

H&P's conclusions have far-reaching repercussions for theories of concept representation and category learning, because most theories assume there *is* a link between intension and extension. Hence, the first aim of this paper is to conceptually replicate their findings. An important difference from H&P's experiments is that we employed a category judgment task to measure category extensions, and a property generation task (Study 1) and property judgment task (Study 2) to measure category intensions. Besides testing whether H&P's findings generalize to other, related measures, we used category judgments instead of typicality ratings to address the concern raised by H&P that different properties may determine category membership as opposed to typicality judgments (for some concepts). Indeed, they speculated that "judgments of feature importance might reflect involvement in category membership decisions rather than typicality, so that variability in the two measures would not match up" (p. 507). The present set-up allows us to test this possibility.

A second goal is to compare H&P's approach with a more direct method to link people's intensions and extensions. Indeed, their (controversial) conclusion might find its roots in the nature of the methodology they used. To address this (potential) concern, we used a more straightforward measure of the extension-intension relation inspired by Hampton (1979). More concretely, we will examine whether a participant's own properties predict her category judgments better than the properties of another participant.

Study 1

Method

Participants. Sixteen adults (8 females); ranging in age from 20 to 46 years old (M_{age} = 26.83) performed both a property generation task and a category judgment task. Another group of 16 adults (10 females), ranging in age from 20 to 55 years old (M_{age} = 30.42) performed a property applicability judgment task.

Materials. The stimuli were sets of 15 possible exemplars from each of 8 semantic categories. Inspired by Ameel et al. (2008), the categories were chosen in pairs of a superordinate (high) level and a corresponding basic (low) level category (specifically, clothes-trousers; fruit-berries; musical instruments-guitars; vehicles-bicycles). The exemplars were presented in the form of pictures in both the category membership judgment and property applicability judgment tasks, whereas in the property generation task only the category names were presented. Within each set of 15 items, ten were presumed category members and five were presumed non-members (based on discussions of the selected materials by two of the authors). Each picture was printed in color on a 11x10 cm cardboard form. Figure 1 displays some of the stimuli (see Appendix for all items). Note that there are two primary differences with H&P's study in terms of the materials used: H&P only selected superordinate categories, some of which were of a non-physical nature (i.e., sports and science) and the exemplars were presented in the form of words instead of pictures. However, these variations from H&P's study were not theoretically motivated.

INSERT FIGURE 1 HERE

Procedure. In a first phase of the study, participants completed the property generation task and then continued to the category membership judgment task, within a single session. The task order assured that participants' generated properties were not influenced by the pictures. That is, participants only saw the pictures in the category membership judgment task. In a second phase of the study, a different group of participants performed the property applicability judgment task, which involved the properties gathered in the first phase of the experiment (i.e., in the property generation task).

In the property generation task, participants were given a MS Excel file that contained eight worksheets, one for each category name. Participants were asked to imagine they had to explain the terms to someone who did not know their meaning. They performed the task individually by typing in the properties. They were instructed to finish one category before moving on to the next one. The categories were presented in different random orders to each participant with superordinate and basic category pairs (e.g., clothes-trousers) never occurring immediately one after the other.

After completing the property generation task, participants were given a link to an online survey (i.e., the category membership judgment task) where each set of 15 pictures was presented and they were asked to click on the pictures of exemplars they judged to be members of the category mentioned above the picture set. Each category name was embedded in a question, for instance, “which pictures below are members of the category fruit?” The category name was written in bold and underlined. All 15 pictures were presented in three rows of five, so that participants could see all the pictures on their computer screen at the same time. Participants were also always able to see the target category name when they were selecting pictures. Before moving on to the next category, participants were asked to check whether they were sure of their answers. If they did, they were allowed to click the ‘next’ button to continue to the next category. In the survey, each participant received a different random order of the categories and pictures.

To select the properties for the property applicability judgment task, we followed the procedure described in McRae, de Sa, and Seidenberg (1997). First, all generated properties were simply tallied for each category name. Synonym properties (i.e., properties that have essentially the same meaning, e.g., *to produce music* and *to make music*) were given an identical code. Properties phrased with an adjective-noun combination (e.g., *heavy iron*) and conjunctive properties (e.g., *red and small*) were split and treated as separate properties if they provided different information. Redundant quantifiers (e.g., *most of them*) were dropped and properties which only mention exemplars of the category (e.g., *apple* for the category *fruit*) were eliminated. The total number of properties (i.e., the number of types, not tokens) generated across participants for each category ranged from 39 to 53 (see Table 1 for the average number of tokens of generated properties per category). In a next step, the 15 possible exemplars per category were combined with the generated category properties to form property by exemplar matrices. Thus, every matrix consisted of 15 columns, one for every exemplar, and 39 to 53 rows, one for every generated property.

INSERT TABLE 1 HERE

In the property applicability judgment task, a different group of participants individually received an excel file in which they indicated whether the exemplars possessed the properties by entering a 1 if the property applied to the exemplar, or a 0 if not. Each participant was randomly assigned to fill in the matrices for two categories. In total, four participants were assigned to each category. All the tasks were conducted in Dutch and none of the tasks had a time limit.

Results

The results are structured as follows. First, we will describe the intensional and extensional measures and explain how these were transformed into similarity matrices. Then, we will test whether similarity between individuals for extensional judgments maps onto similarity between individuals for intensional judgments by correlating extension and intension similarity matrices (i.e., H&P’s approach). Finally, we will introduce a new method using an individual’s properties to predict her own category judgments and compare the results.

Extension similarity matrix. The category judgments were first quantified by scoring each decision as 0 or 1, depending on whether the item was judged a non-member or a member of the category (see Table 1 for the mean and standard deviation of the category judgment scores, the proportion of yes responses, per category). These scores were then tabulated for each participant in each category. To measure between-participant consistency, Cronbach’s alpha reliability coefficients (Lord, Novick, & Birnbaum, 1968) were calculated for the category judgment scores. Agreement between people (and a large sample size) translates into a high alpha. The reliability coefficients obtained in the present study varied between .92 and .99 across the eight categories. For each category, a participant by participant (16x16) similarity matrix was then constructed with the correlations between participants’ category judgments. This matrix shows the similarity between all pairs of participants in their category judgments.

Intension similarity matrices. Based on the property generation task, we constructed two intension similarity matrices. The first one was a property overlap measure and the second one was derived from the property applicability scores.

For the first matrix, property overlap scores were computed (see Tversky, 1977). That is, for every two participants and every category, the number of common properties (i.e., properties that were generated by both participants) was divided by the sum of common and distinctive properties (i.e., the number of unique properties from both participants). For each category, we then constructed a 16x16 similarity matrix with participants’ property overlap scores. This matrix will be termed the “property overlap similarity matrix”.

For the second matrix, summed property applicability scores were computed for each exemplar separately for each participant based on the properties that she herself generated. The idea is that if an exemplar possesses many of the properties generated by person X (i.e., the exemplar has a high *property applicability score*), it is more likely to be included as a category member by that person. The procedure to calculate the property applicability scores is illustrated in Figure 2 for the category fruit (the same holds for the other seven categories). The applicability judgments (0 or 1) for each property × exemplar combination were first summed over the four participants who completed the property applicability judgment task, resulting in *property applicability scores* that ranged from 0 to 4. The Cronbach's alpha reliability coefficient for these property applicability judgments varied between .78 and .90 across the eight categories. Using the specific properties a participant generated (i.e., *individual* properties), summed property applicability scores were then calculated by adding the *property applicability scores* of the individual properties for each of the 15 exemplars separately. For instance, to calculate the summed property applicability score of the exemplar *banana* for participant X, the applicability scores for *banana* were summed across all the properties generated by participant X (see the vertical box under the exemplar *banana* in Figure 2). This procedure was carried out for all the 15 exemplars separately and for every participant using her own individual properties. The result is a vector with 15 elements for each participant × category combination, representing the degree to which each exemplar possesses the properties generated by that participant. Finally, a 16x16 similarity matrix for participants was constructed using the correlations between participants' summed property applicability scores. To avoid confusion, this matrix will be termed the "property applicability similarity matrix". Thus, both the property applicability and the property overlap similarity matrices provide some insight into how similar or dissimilar participants are in terms of their category intensions[3].

INSERT FIGURE 2 HERE

Correlation between similarity matrices. In order to discover whether there is a link between extension and intension, we correlated, for each category, the lower triangular extension similarity matrix with the corresponding property applicability and property overlap similarity matrices. Before extension and intension similarity matrices were correlated, the central tendency and variability of the similarities was checked. Table 2 shows the average of each of the three similarity measures per category, whereas Figure 3 shows the distribution of these measures across all eight categories. Because our data were non-normally distributed (with skewness of the similarities from the category judgment, property overlap, and property applicability matrices: -1.93, 1.00, and -1.17, respectively), we used Spearman's non-parametric rank-order correlation to examine the correspondence between the extension and intension matrices[4].

INSERT TABLE 2 HERE

INSERT FIGURE 3 HERE

----

[3] One could argue that property applicability similarity is an imperfect reflection of intensional similarity. Two people might generate completely different properties, yet if these yield similar amounts of evidence for a given set of exemplars belonging in the category, the property applicability scores could be strongly correlated. It should be noted that property applicability similarity does correlate reliably with property overlap similarity, as will be shown later.

[4] All correlations used to construct similarity matrices are Pearson product-moment correlations.

For each category, we first calculated the correlations between the extension similarity matrix and the two different intension matrices (see the first and third columns of Table 3 under the subheading “Within-category”). Collapsing across categories, the average correlations were close to zero: *M* = .03 for property applicability and *M* = -.07 for property overlap. Following H&P, we also correlated the extension similarity matrix from a particular category (e.g., *clothes*) with the intension matrices from the other seven categories (see columns with subheading “Between-category”). This procedure provides a control for non-specific similarities in how people may be approaching each task. Mean correlations were very close to zero. Using independent samples *t* tests to compare the (Fisher’s Z transformed) eight within- and 56 between-category correlations, we found no significant difference for category judgment – property applicability (*p* = .97) and category judgment - property overlap (*p* = .18). These results conceptually replicate H&P’s findings and might thus suggest that there is no clear link between people’s individual category extension and intension.

For completeness sake, Table 3 also shows the correlations between the two different intensional matrices. Seven of the 8 categories showed significant positive correlations, with a mean of .41. This result is to be expected, given how the matrices were constructed.

INSERT TABLE 3 HERE

Predicting category judgments from individual properties. The previous analyses showed correlations between intension similarity matrices and extension similarity matrices that were close to zero. This *could* mean that participants’ category judgments were not based on their own properties. It is, however, possible that properties from other participants contain useful information for the prediction of one’s own category judgment. In other words, it is feasible that people use properties in judging category membership that they do not come up with in a property generation task (Bellezza, 1984), but that are nevertheless generated by other participants. Thus, a mixed effects logistic regression analysis was run to investigate to what extent the particular properties that a person generated (i.e., the *individual* properties) contribute to his/her own particular category judgment.

The analyses were carried out in R (version 3.1.2) using the lme4 package (Bates, Maechler, Bolker, & Walker, 2014). Category judgment of a given individual to a given exemplar, a binary variable, was included as the response of interest and two fixed effects were included. The first one comprised the individual property applicability scores (i.e., based on the set of properties that that *specific* participant generated, see Figure 2) and the second predictor contained the *residual* property applicability scores (i.e., based on the properties that were generated by the *other* participants, see Figure 2). For each participant × category combination, we z-transformed both the individual and residual property applicability scores.

In addition, category level (basic and superordinate levels of a category) and domain (clothes, fruit, musical instruments, and vehicles) were included as dummy-coded covariates (these effects are not the main interest of the analyses). Random effects for participants and items (i.e., the 120 different pictures) were also included. Following the suggestion from Barr, Levy, Scheepers, and Tily (2013), the random effects structure was maximal except when it concerned the control variables category level and domain, and random correlations were excluded as well[5]. The analysis revealed that *individual* properties have a significant contribution to the prediction of the person-specific category judgments ($\beta = 0.87$, $SE = 0.16$, $\chi^2(1) = 20.20$, $p < .001$). This might lead to the conclusion that people's own properties are directly linked to their category extension. We also found a significant effect of the residual properties ($\beta = 2.02$, $SE = 0.26$, $\chi^2(1) = 50.70$, $p < .001$). This means that properties that participants did not generate also play a role in predicting category judgment.

[5] The analysis code can be found on Open Science Framework (https://osf.io/8vewz/?view_only=1b287bc02fab49fdbbbfe923332f1cd0).

However, if *all* properties are (to some extent) predictive for category judgments, it shouldn't be a surprise that we found a significant effect of the individual property applicability scores in the previous analyses. Indeed, these results did not prove that one's own properties are *special*. Thus, in an additional analysis, we sought to compare how well a person's properties predict their own category judgments as opposed to other people's category judgments. To examine whether intension predicts extension at the individual-specific level, we conducted a similar analysis, except that we now shuffled the category judgments of all participants. That is, each participant's individual property applicability score was paired with another participant's category judgments. For example, the category judgment scores of Participant 1 were paired with the individual property applicability scores of Participants 2, 3, 4,…., or 16 (there were 16 participants in total – see Table 4 for a simplified illustration). In the previous analyses a person's category judgments were predicted by the specific properties she generated and the properties generated by the remaining 15 participants. Now, a person's category judgments will be predicted by a different person's properties and the 15 other participants' properties. The latter entails that a person's own properties are now included in the residual property applicability scores.

INSERT TABLE 4 HERE

A similar mixed effects logistic regression analysis was again run, but this time, instead of using category judgment as the dependent variable, we used the shuffled category judgments data as the response of interest. We repeated this procedure for 1,000 random shuffles of the category judgment data. Each time we compared the regression weights obtained from this model with the regression weight obtained from the previous analysis (i.e., using the original *non*-shuffled category judgment data). If there is a link between someone's category intension and extension, we would expect that the regression weight of the *individual* property applicability scores would be higher in the original analysis compared with the shuffled data. In contrast, we expect that the regression weight of the *residual* property applicability scores would be higher with the shuffled data, because a participant's own properties are now actually included in the calculation of the residual property applicability scores. We found that for 98.69% of all the simulations, as expected the original regression weight of *individual* property applicability scores was higher than the one obtained using the shuffled data. We also found that in 97.78% of all the simulations, the regression weight of *residual* property applicability scores was lower than the one obtained using the shuffled data[6]. Taken together, these results do provide evidence that there is a relation between people's category extensions and their intensions, for certain categories at least.

Study 2

In order to replicate and extend the findings of Study 1, Study 2 was conducted using a larger sample (i.e., 80 participants and 24 exemplars per category). The same methods, correlating extension and intension similarity matrices and mixed effects logistic regression analyses, were again used to examine the extension-intension relation. There were a few differences compared to Study 1. First of all, we employed a property judgment task instead of a property generation task to measure category intension, and hence the calculation of the property applicability scores was slightly different from Study 1. The reason was that, during a property generation task, people may forget to mention certain properties or give properties they are not actually using when judging category membership. Although their own properties may still predict their category judgments to some degree, there is ample room for improvement. The latter statement is supported by the finding in Study 1 that residual properties (i.e., properties generated by *other* individuals) were predictive for person-specific

[6] In some cases the shuffling caused convergence problems during the mixed effects logistic regression analysis. These were removed from the analyses.

category judgments too. In other words, the few properties a person generates play a role in their category judgments, but properties they do not generate are an even bigger factor. The notion that people fail to generate some crucial properties may make it difficult to obtain a reliable intension similarity matrix. As found in Bellezza's (1984) study, the test-retest reliability of generating properties for category terms can be quite low. He argued that since there is no well-defined meaning of a word, it is difficult to retrieve the same information on different occasions. This could in turn explain why we did not find a correlation between extension and intension similarity matrices.

So, in order to address the concern that people cannot consciously access all relevant properties that they employ to define a concept, we used a property *judgment* task instead of a property *generation* task. If people tacitly "know" which properties to use when making a category decision, they may be able to recognize them in a property judgments task even though they (partly) fail to retrieve them during a property generation task.

A second difference with respect to Study 1 is that we now added a filler task (i.e., solving analogies), between the property judgment and category judgment task, in order to eliminate or at least reduce any potential carry-over effect. Finally, we selected three different categories for this study, from Verheyen and Storms (2013), which were also used by H&P (as opposed to the eight categories used in Study 1): insects, tools, and sciences. The 24 items per category, including clear members, clear non-members, and borderline cases, were also taken from Verheyen and Storms as were the property applicability matrices. This study was pre-registered on Open Science Framework (OSF, https://osf.io/bqekx/?view_only=15f087a23f8f4a7bbc417c7b030c7c5f), and all data and analysis code, can be found using this link (https://osf.io/8vewz/?view_only=1b287bc02fab49fdbbbfe923332f1cd0).

Method

Participants. Eighty participants (50 females), ranging in age from 18 to 32 years old (M_{age} = 19.05) performed a property judgment task and a category judgment task. They participated voluntarily or received study credit for their participation. Five participants unexpectedly showed no variability in their property or category judgments for at least one complete category, so they were excluded from the analyses[7].

Materials. As in H&P, the materials were taken from Verheyen and Storms' (2013) study. In the latter, eight categories were used, representing four different category types (animals: fish and insects; artifacts: tools and furniture; activities: sports and sciences; and borderline artifact-natural-kind categories: fruit and vegetables). To keep the task practically feasible for participants, we reduced the number of categories to three, based on these criteria: (1) we wanted to have one category from each type, except for the artifact-natural-kind group, because the category fruit was already included in Study 1; (2) since the main aim of our study is to investigate consistency between extension and intension, inter-individual diversity is important. To determine this, we used the category judgment data from Verheyen and Storms (2013). After case-wise removal of missing data and exclusion of participants without any variability in their category judgments, we calculated Cronbach's alpha and average pair-wise correlations (i.e., correlations between subject X's category judgments and subject Y's category judgments). The following categories demonstrated the most inter-individual variability (from high to low): sciences, insects, sports, and tools. As we wanted one category per type, we ultimately selected sciences (activities), insects (animals), and tools (artifacts).

Verheyen and Storms selected 24 items per category, comprising clear members, clear non-members and borderline cases, all of which were used in the category judgment task. In

[7] We did not take into account the possibility that some participants would give the same response to all questions, which is why this exclusion criterion was not mentioned in the pre-registration plan. Showing no variance across items rendered it impossible to compute correlations with other participants.

addition, they gathered property generation data as well as property applicability judgments for all categories. All resulting properties (i.e., 39 for sciences, 35 for insects, and 34 for tools) were included in the property judgment task. In contrast to Study 1, exemplars were presented as words instead of pictures.

Procedure. Each participant was given a link to an online survey consisting of three tasks presented in the following order: a property judgment task, an analogy test (i.e., a filler task), and a category judgment task. Participants were tested in one session.

In the property judgment task, participants were shown a list of properties underneath a category name. They were asked to judge whether each property was true for that category name by clicking a “yes” button or a “no” button. They could only continue to the next category if they had given a response to all the properties. The order of the categories and the properties within a category were randomized for each participant.

A similar procedure was used in the category judgment task. However, instead of a list of properties, participants were shown a list of exemplars (presented as words) and they had to judge category membership of each exemplar by clicking a “yes” button if they thought that the exemplar belonged to that category or a “no” button if they thought it wasn’t a member of the category. They had to give a response to all 24 exemplars before they could go on to the next category. Each participant received a different random order of categories and exemplars within a category.

The analogy test consisted of 10 multiple choice questions such as “wrist : elbow :: ankle : ? ” (where the correct response was knee in this case). All tasks were conducted in Dutch and none of the tasks had a time limit.

Results

Extension similarity matrix. Again, category judgments were quantified by giving a score of 0 or 1, based on whether the item was judged a non-member or a member of the category. Table 5 shows the mean and standard deviation of the category judgments scores (the proportion of yes responses) per category. Cronbach's alpha reliability coefficients for the category judgment scores were .99, .99, and .98 for the categories insects, tools, and sciences, respectively. A participant by participant (75x75) similarity matrix was then constructed for each category with the correlations between participants' category judgments.

INSERT TABLE 5 HERE

Intension similarity matrices. Two intension matrices were constructed, the first one was based on property judgments and the second one was based on property applicability scores. To construct the first matrix, the property judgments were quantified by giving a score of 1 if the property was judged to apply to the category and a score of 0 if not (see Table 5 for the mean and standard deviation of the property judgments, expressed as the proportion of properties judged to be true of the category). These scores were then tabulated for each participant in each category. Cronbach's alpha reliability coefficients for the property judgment scores were .96, .97, and .97 for the categories insects, tools, and sciences, respectively. A participant by participant (75x75) similarity matrix was constructed for each category by correlating participants' property judgment scores, which we will call the "property judgment similarity matrix".

To construct the second intension matrix, a similar procedure as in Study 1 was employed to compute the *individual summed property applicability scores* (see Figure 4), only this time, the *property scores* were taken from Verheyen and Storms (2013). Unlike Study1, the *individual properties* were those properties that were considered to apply to a category according to a particular participant. Using these *individual properties,* summed property applicability scores were then calculated by adding the *property applicability scores* for the properties of each of the 24 exemplars separately. A 75x75 similarity matrix was constructed with the correlations between participants' summed property applicability scores. This matrix will be termed the "property applicability similarity matrix".

INSERT FIGURE 4 HERE

Correlation between similarity matrices. As in Study 1, the similarity matrices from each task and each category were correlated. Before doing so, the central tendency and variability of the similarities was first checked. Table 6 shows the average of each of the three similarity measures per category, whereas Figure 5 shows the distribution of these measures across all categories. Spearman's non-parametric rank-order correlations were again used to measure the relation between extension and intension, since our data were non-normally distributed (with skewness of the similarities from the category judgment, property judgment, and property applicability matrices: -0.19, -0.12, and -2.81, respectively).

INSERT TABLE 6 HERE

INSERT FIGURE 5 HERE

Table 7 shows the correlations between the extension similarity matrix and the two different intension matrices per category. Collapsing across categories, the mean correlations were again found to be close to zero (see the first and third columns of Table 7 under the subheading “Within-category”): *M* = .03 for property judgments and *M* = .06 for property applicability[8]. These results, again, confirm H&P’s findings, suggesting that there is no (strong) link between a person’s category extension and his/her intension.

INSERT TABLE 7 HERE

Predicting category judgments from individual properties. A mixed effects logistic regression analysis was again run to investigate to what extent the properties that a person considers applicable to a category (i.e., the *individual* properties) can predict his/her own category judgment. Analogous models as in Study 1 were used in the analyses. Category judgment was the response of interest and two fixed effects were included. The first one consists of the *individual* property applicability scores (i.e., based on the properties that apply to a category according to a participant), and the second one contained the *residual* property applicability scores (i.e., based on the properties that were *not* endorsed by a given participant). These can be considered residual properties because they *were* responses from participants in Verheyen and Storms’ (2013) property generation task. For each participant × category combination, we once again z-transformed both the individual and residual property applicability scores. In addition, category (with levels insects, tools, and sciences) was included as a covariate and the same random effect structure as in Study 1 was used.

The analysis revealed that *individual* properties contribute significantly to the prediction of the person-specific category judgments ($\beta = 2.11$, $SE = 0.22$, $\chi^2(1) = 67.00$, $p < .001$), whereas the *residual* properties did not ($p = .30$). These findings seem to support the

[8] Because people presumably endorsed properties that they may have failed to come up with during the property generation task (Study 1), the resulting applicability scores are more similar. However, the results showed comparable correlations with the other measures of intensional similarity (.41 in Study 1 and .37 in Study 2), mitigating concerns about potential ceiling effects. The high values should not come as a surprise, given that people generally agree on category extension (see Verheyen & Storms, 2013 and the high Cronbach's alpha for category judgment).

conclusion that people’s set of properties are directly linked to their category extension. They also suggest that, when presented with potential properties, participants are able to select those properties that drive their category membership judgments. In contrast, when people have to generate properties themselves, as was the case in Study 1, they only come up with a subset of all properties that they actually take into account when making category membership decisions. That is probably why residual properties significantly predicted category judgments in Study 1, but not in the present study.

Finally, the same shuffling procedure as in Study 1 (see Table 4) was run to compare how well a person’s intension predicts her own category judgments as opposed to other people’s category judgments. The results showed that the original regression weight of the *individual* property applicability scores was higher than the one obtained using the shuffled data in 100% of the simulations. Similarly, the regression weight of the *residual* property applicability scores was lower than the one obtained using the shuffled data in 99.90% of the simulations. These results strongly suggest that there is a relation between people’s category extensions and their intensions.

General Discussion

Across two studies, we investigated the relationship between category extension and intension in eleven semantic categories. It is often tacitly assumed that there is a (strong) extension-intension link, yet, a recent study by H&P called this hypothesis into question. They found that systematic inter-individual variability in extensional beliefs did not significantly correlate with inter-individual variability in intensional beliefs.

Because of the theoretical importance of these findings, we sought to extend H&P’s findings using other measures of category extension and intension. To capture category intension we asked participants to describe categories based on their own perspective (Study 1) or to judge whether a set of properties are true for a particular category (Study 2). In addition, instead of typicality judgments, we used a category judgment task to measure individuals' category extensions. Although typicality judgments are a common and valid measure of category extension, it is a gradual measurement that allows members of a category to vary in how good they are as an example of a category or how typical they are of the category. On the other hand, category judgment is a more decisive binary measurement, which might capture different information. More specifically, H&P raised the possibility that different properties may determine category membership as opposed to typicality. So to rule out the possibility that H&P's findings were merely caused by the use of typicality ratings, we employed category judgments as a measure of extension.

Using H&P's method (i.e., correlating extension and intension similarity matrices) in both studies, we found evidence suggesting that similarity between individuals for extensional judgments did not map onto similarity between individuals for intensional judgments. These findings indicate that H&P's results were not merely a product of the particular intensional and extensional measures they used. So contrary to popular belief, it may appear that there is a disconnection between category extensions and intensions. However, H&P's approach is a bit unconventional as it does not directly compare an individual's intension with her extension. As we will discuss later, the nature of this procedure brings about some methodological concerns that may invalidate their conclusions.

A match made in heaven after all?

In a follow-up analysis, we directly related people's category judgments to the properties they themselves generated or endorsed for a certain category. The latter analysis, however, provided evidence *in favor* of a link between intension and extension. That is, the properties a person generated (Study 1) or endorsed (Study 2) were generally a better predictor of her category judgments than of the category judgments of other people. The question is, why do both approaches, using the same dataset, lead to conflicting findings? And what should we actually conclude about the relation between a person's category intension and his/her extension?

We see three possible explanations for these differences. First, even though the reliability coefficients for the property applicability judgments were high (e.g., in Study 1, they varied between 78 and .90), there seems to be some disagreement among participants as to which properties apply to which exemplars. This could be an issue for H&P's method, because two people with the same category intension, may disagree about which exemplars possess certain properties resulting in (partly) different category extensions (and vice versa). Thus, intensional similarity does not necessarily translate into extensional similarity. Note that every participant with "atypical" property applicability views can influence *N* - 1 data points in the property applicability similarity matrix (i.e., all similarities with the other participants). This could be one of the reasons why both H&P and the present study showed non-significant correlations between intensional and extension similarity.

On the other hand, the mixed effects logistic regression method might not be affected by idiosyncratic property applicability views to the same extent. If a person holds unconventional beliefs about which exemplars possess which property, her individual property applicability scores will be inaccurate. Whereas the *appropriate,* yet unknown, property applicability scores of that participant may predict her category judgments, the *derived* applicability scores may not. This would in turn decrease the fit of the mixed effects logistic regression model, but the relation between the derived applicability scores and the category judgments will still hold for the average person. Hence, the regression weight of the person-specific properties will differ significantly from zero, as was the case in both Study 1 and 2. Furthermore, because the subsequent shuffling procedure considers all participants simultaneously, atypical participants will literally get lost in the shuffle. That is to say, if the category judgments are shuffled, nothing changes for participants whose derived applicability scores are an imperfect reflection of their true applicability scores: on average, the derived applicability scores will predict *someone else's* category judgments equally badly compared to their *own* category judgments (or even slightly worse). The story is different for the other participants (the majority in all likelihood), whose derived applicability scores more accurately reflect their true applicability scores. On average, their derived applicability scores will predict someone else's category judgments considerably worse than their own, assuming of course that there is an intension-extension link and inter-individual variability in both measures. So that is probably why, as a whole, individual property applicability scores are a better predictor of one's own category judgments as opposed to someone else's judgments. More generally, because this procedure considers all participants at the same time, it is probably more likely to detect subtle effects.

A second reason why the two approaches may lead to different conclusions has to do with the reliability of the similarity matrices. More precisely, the correlation between the extension and intension similarity matrices is necessarily constrained by the reliability of the two matrices. Our data do not allow us to estimate the reliability of the similarity matrices (since, unlike H&P, our participants performed every task just once), but in H&P's data, the reliability estimates were rather low (ranging from .18 to .51). If one takes this unreliability into account, for instance by applying Spearman's correction for attenuation formula, the resulting correlations are considerably higher. On average, the resulting corrected correlation was .35, which seems to suggest that there might be a (weak) relation between extension and intension after all. On the other hand, H&P's meta-analysis of their effect sizes suggested a 95% CI for the correlation of between +0.1 and -0.1 which is a very small effect.

Finally, based on the literature (Ameel et al., 2008; Malt et al., 1999), there is arguably no *perfect* link between extension and intension. This notion likely compounds the reliability issue. If on the one hand, the extension and intension similarities are not very reliable, and on the other hand, the link between extension and intension is, at the very least, imperfect, one may have trouble finding significant correlations between the intension and extension similarity matrices. Furthermore, there may be inter-individual variability in the extension-intension link. Put differently, some people may be very consistent in their category extensions and intensions, whereas other people may have a weaker link between their extensions and intensions. In conclusion, the present study confirms H&P's results in that extensional similarity does not necessarily map onto intensional similarity. Taken at face value, they may challenge the long-held belief that there is a direct, but perhaps imperfect, relation between category intension and extension. However, the method of correlating intensional and extensional similarity matrices is a rather indirect way to test whether intensions and extensions are connected. The outcome critically depends on a) agreement about which properties apply to which exemplars, b) the reliability of both similarity matrices, c) the strength of the extension-intension link, and d) inter-individual variability in the strength of the extension-intension link. The combination of these factors might result in very low and even negative correlations (the latter due to random noise). In contrast, directly predicting a person's category judgments from the properties she generated or endorsed does not suffer (to the same extent) from these issues. Using this, presumably more sensitive method, we did find a significant relation between a person's category intension and his/her extension, indicating that properties are important even though they might not tell the whole story (Ameel et al., 2008; Malt et al., 1999).

Acknowledgments

FMD gathered and analyzed the data presented in this paper, and participated in the writing of the manuscript. TH contributed in the writing of the manuscript and analyzed the data. All four authors discussed the findings thoroughly, read, and approved the final version of the manuscript. The research described in the manuscript was sponsored by grants DBOF/12/010 and OT/10/024 from the Leuven Research Council. We are grateful to Lieselot Verlooy, Liesbeth Berckmans, and Sarah Casaer for assisting in the data collection and Eef Ameel for her help in setting up the study and her valuable suggestions. TH is a postdoctoral fellow of the Research Foundation-Flanders (FWO-Vlaanderen). Correspondence should be addressed to FMD, Department of Psychology, University of Leuven, Belgium. Email address: Farahmutiasari.Djalal@kuleuven.be.

References

Ameel, E., Malt, B.C., & Storms, G. (2008). Object naming and later lexical development: From baby bottle to beer bottle. *Journal of Memory and Language*, *58*(2), 262-285. doi:10.1016/j.jml.2007.01.006

Ameel, E., Malt, B.C., & Storms, G. (2014). Steps along a continuum of word knowledge: Later lexical development through the lens of receptive judgments. *Language Learning and Development, 10*(3), 234-262. doi:10.1080/15475441.2013.840485

Aristotle (1961). *De anima, Books II and III,* trans. D.W. Hamlyn. Oxford University Press. (Original work published in 4th century B.C.).

Barr, D.J., Levy, R., Scheepers, C., & Tily, H.J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language,* *68*(3), 255-278. doi:10.1016/j.jml.2012.11.001

Bates, D., Maechler, M., Bolker, B. & Walker, S. (2014). lme4: Linear mixed-effects models using Eigen and S4 (Version 1.1-7). Retrieved from: http://CRAN.R-project.org/package=lme4

Bellezza, F.S. (1984). Reliability of retrieval from semantic memory: Noun meanings. *Bulletin of the Psychonomic Society, 22*(5), 377-380. doi:10.3758/BF03333850

Caplan, L.J., & Barr, R.A. (1989). On the relationship between category intensions and extensions in children. *Journal of Experimental Child Psychology, 47*(3), 413-429. doi:10.1016/0022-0965(89)90022-2

Frege, G. (1948). Sense and reference. *The philosophical Review, 57*(3), 209-230. doi:10.2307/2181485

Hampton, J.A. (1979). Polymorphous concepts in semantic memory. *Journal of Verbal Learning and Verbal Behavior, 18*(4), 441-461. doi:10.1016/S0022-5371(79)90246-9

Hampton, J.A., Dubois, D., & Yeh, W. (2006). Effects of classification context on categorization in natural categories. *Memory & Cognition, 34*(7), 1431-1443. doi:10.3758/BF03195908

Hampton, J.A., & Passanisi, A. (2016). When intensions do not map onto extensions: Individual differences in conceptualization. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 42*(4), 505-523. doi:10.1037/xlm0000198

Heit, E., & Barsalou, L.W. (1996). The Instantiation Principle in Natural Categories. *Memory, 4*(4), 413-451. doi:10.1080/096582196388915

Lord, F.M., Novick, M.R., & Birnbaum, A. (1968). *Statistical theories of mental test scores*. Oxford, England: Addison-Wesley.

Malt, B.C., Sloman, S.A., Gennari, S., Shi, M., & Wang, Y. (1999). Knowing versus naming: Similarity and the linguistic categorization of artifacts. *Journal of Memory and Language, 40*(2), 230-262. doi:10.1006/jmla.1998.2593

McRae, K., de Sa, V.R., & Seidenberg, M.S. (1997). On the nature and scope of featural representations of word meaning. *Journal of Experimental Psychology: General, 126*(2), 99-130. doi:10.1037/0096-3445.126.2.99

Murphy, G.L. (2002). *The big book of concepts*. Cambridge: MIT press.

Nosofsky, R.M. (1984). Choice, similarity, and the context theory of classification. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 10*(1), 104-114. doi:10.1037//0278-7393.10.1.104

Rosch, E. (1975). Cognitive representations of semantic categories. *Journal of Experimental Psychology: General, 104*(3), 192-233. doi:10.1037/0096-3445.104.3.192

Rosch, E., & Mervis, C.B. (1975). Family resemblances: Studies in the internal structure of categories. *Cognitive Psychology, 7*(4), 573-605. doi:10.1016/0010-0285(75)90024-9

Smith, E.E., & Medin, D.L. (1981). *Categories and concepts*. Cambridge: Harvard University Press.

Smith, E.E., Shoben, E.J., & Rips, L.J. (1974). Structure and process in semantic memory: A featural model for semantic decisions. *Psychological Review, 81*(3), 214-241. doi:10.1037/h0036351

Spalding, T.L., & Gagné, C.L. (2013). Concepts in Aristotle and Aquinas: Implications for current theoretical approaches. J*ournal of Theoretical and Philosophical Psychology, 33*(2), 71-89. doi: 10.1037/a0029990

Storms, G., De Boeck, P., & Ruts, W. (2000). Prototype and exemplar-based information in natural language categories. *Journal of Memory and Language, 42*(1), 51-73. doi:10.1006/jmla.1999.2669

Tversky, A. (1977). Features of similarity. *Psychological Review, 84*(4), 327-352. doi:10.1037/0033-295X.84.4.327

Verheyen, S., De Deyne, S., Dry, M. J., & Storms, G. (2011). Uncovering contrast categories in categorization with a probabilistic threshold model. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 37*(6), 1515-1531. doi:10.1037/a0024431

Verheyen, S., & Storms, G. (2013). A mixture approach to vagueness and ambiguity. *PloS one, 8*(5), e63507. doi:10.1371/journal.pone.0063507