Mem Cogn (2011) 39:1290–1300
DOI 10.3758/s13421-011-0102-3

On domain differences in categorization and context variety

Steven Verheyen · Daniel Heussen · Gert Storms

Published online: 3 May 2011

Abstract Membership in many natural categories is considered all-or-none, while membership in most artifact categories is found to be graded. We introduce an alternative for the prevailing view that this domain difference in categorization results from representational differences. The context variety account posits that an item's gradedness reflects the variety of contexts it appears in. Items that feature in a variety of contexts are assumed to be more likely to elicit a graded categorization response, since the suggested target category only provides one of many solutions to the question of the item's identity. We review earlier work that suggested a domain difference in context variety, with artifactual items appearing in a greater variety of contexts than natural ones. The context variety domain difference is established in two separate experiments but is shown not to explain the domain difference in categorization. A selection of artifactual and natural items, for which the domain difference in context variety is reversed, is presented for categorization in a third experiment. This selection, too, fails to provide evidence for the context variety account of categorization differences. The domain difference in categorization is shown to be robust against this manipulation. Context variety appears to have no bearing on categorization, so the context variety account is not a sustainable alternative to accounts that posit representational differences between natural and artifact categories.

Keywords Context variety · Gradedness · Categorization · Typicality · Artifacts · Natural kinds · Essentialism

S. Verheyen (✉) · D. Heussen · G. Storms
Department of Psychology, University of Leuven,
Tiensestraat 102, Box 3721,
BE-3000 Leuven, Belgium
e-mail: steven.verheyen@psy.kuleuven.be

Natural and artifactual stimuli that are matched for typicality in their respective target categories present with a difference in the manner in which they are categorized. While the natural items (e.g., avocado, pumpkin, sage) are generally categorized in an absolute manner, the artifactual items (e.g., funnel, knife, wheelchair) tend to be categorized in a more continuous manner. Participants in Diesendruck and Gelman (1999), for instance, were more likely to judge natural items as definitely members or definitely not members of their respective categories than to make such absolute judgments of artifactual items of equal typicality. The artifact instances tended instead to be awarded an intermediate or graded membership response. This phenomenon has been replicated ever since, using a number of different methodologies in various populations (see, e.g., Estes, 2003, 2004; Rhodes & Gelman, 2009). Its implications are potentially far reaching. The phenomenon points toward a difference between typicality and categorization and suggests that there are different types of categories with different sorts of representations.

However, notable exceptions place constraints on any account that explains the domain difference in categorization by positing representational differences between natural and artifact categories. A study by Kalish (2002) does not support a strong divide between "absolute" natural and "graded" artifact categories, for instance. Instead, Kalish (2002) found that within each of the domains there was significant variability in beliefs about category structure, with some artifact categories believed to be as absolute as some natural categories. Estes (2003) conducted a categorization study in which participants could opt for absolute or partial membership judgments. Although his results corroborated previous findings, in that the mean proportion of absolute membership judgments was higher for natural items than for artifacts, certain individual artifact items were found to be as prone to absolute judgments as some of the natural items. Estes (2003) remarked that this variability within domains severely reduces the predictive power of the prevailing representational accounts: While they may nominally account for the gradedness exhibited by categories in both domains, they offer no basis on which to predict whether a given category will have an absolute or a graded structure (p. 208). To illustrate this point, he referred to a version of psychological essentialism on the part of Gelman and Hirschfeld (1999), which states that the extent to which both natural and artifactual items possess a category's essence determines the gradedness that they display.

The aim of the present work is to test an alternative, predictive account of the categorization differences found between and within the domains of natural items and artifacts. We will review a series of studies that can be taken to support the idea that natural items and artifacts differ with respect to the number of contexts they appear in. (The term *context*, here, refers to the different senses, meanings, and linguistic uses associated with an item.) According to this evidence, natural items are expected to occur in a limited number of contexts, while artifacts would tend to occur in a variety of circumstances. This domain difference in what we will call *context variety* could be said to explain the domain difference in categorization. If participants were to interpret a suggested target category for the artifact items as just one of the many possible manners in which the items could be used, they might be less inclined to award the items absolute membership responses. Graded categorization responses might then reflect the extent to which there are more aspects to the item's identity than are conveyed by the suggested target category. For instance, participants could be responding to the "knife–weapon" pairing in a continuous manner, because a knife has many applications besides that of a weapon. Depending on the circumstances, it makes perfect sense to categorize a knife as a tool, a kitchen utensil, or a piece of tableware. The natural item avocado, on the other hand, is generally only thought of as a fruit or a food ingredient. The same rationale can be applied to account for categorization differences within a domain. These too could be expected to reflect differences in context variety, with items that have few contexts associated with them being less likely to receive partial-membership judgments than are items associated with many contexts. The context variety account of categorization differences thus predicts a positive relationship between context variety and gradedness, both across and within domains.

The customary paradigm for eliciting domain differences in categorization, with its presentation of borderline items for categorization under loose constraints, is particularly prone to a context variety interpretation. In regular discourse, items at the borderline of the category are less likely to be endowed with the corresponding category term than are the prototypical category instances. Dissemination of the category term's meaning within a language group would be too big of a hurdle, otherwise (Hampton, 2007; Malt, 2010). With the pairing of (borderline) item and category only occurring rarely, language users rely on other cues to establish or discard category membership. Normally, the context in which the categorization is to take place would be paramount in resolving the issue (Hampton, Dubois, & Yeh, 2006; Malt & Sloman, 2007). However, categorization tasks that merely list combinations of items and categories lack such contextual constraints. Hampton, Storms, Simmons, and Heussen (2009) had participants categorize items as (1) clear members, (2) intermediate members, or (3) clearly not members of suggested target categories. From among the response alternatives offered to explain intermediate choices, ambiguity was selected 51% of the time for the artifacts and 27% of the time for the natural items. This response alternative stated that intermediate items are inherently ambiguous, and that the precise answer would require more information about the exact context involved.

It has been suggested that in the absence of a clear categorization context, participants might be inclined (1) to come up with a context of their own (Hampton et al., 2006) or (2) to reinterpret the categorization task as requiring a decision on the best possible name for the item (Hampton, Estes, & Simmons 2007). Gradedness might then be higher the more contexts an item appears in, for it becomes less likely (1) that a recalled context will afford the suggested category term or (2) that the single sense suggested by the target category will be the best possible one.

Both the philosophy-of-language literature and the decision-making literature carry support for the context variety interpretation of graded categorization. Counterfactual uses of language terms are an intrinsic part of conceptual role semantics (Block, 1998). In this extension of the use theory of meaning (Wittgenstein, 1953), imagined alternative uses of language terms also affect their perceived meaning. It has long been established that participants spontaneously come up with alternatives for the stimuli they are confronted with (Bear, 1974; Garner, 1966). Many decision-making theories incorporate the assumption that the mental availability of these alternatives influences the decisions that are to be made with regard to these stimuli (e.g., Kahneman & Miller, 1986; Tversky & Koehler, 1994; Windschitl & Wells, 1998). It is therefore plausible that the availability of alternative contexts in which an item could be used influences the decision of whether or not to categorize the item in a graded manner.

Of course, for the context variety account to explain domain differences in categorization, artifacts would have to appear in a greater variety of contexts than natural items. The study by Hampton (2009), described above, already suggested that artifacts rely more heavily on a specified context for categorization than natural items do. Before we start our exposition of the three experiments that were conducted to test the context variety account, we will review other literature in support of a domain difference in context variety.

A domain difference in context variety?

To our knowledge, the hypothesis that artifacts appear in a greater variety of contexts than do natural kinds has yet to be explicitly tested. There does seem to be general agreement that, while a generally accepted delineation of natural items exists, the same is not true of artifacts. Artifacts can participate in a multitude of groupings, whether or not these can be identified by an established label (Keil, Greif, & Kerner, 2007).

In studies in which participants are required to generate exemplars of categories, for instance, some of the same artifacts are generated in response to a number of category labels, while the natural items are generally produced in response to a single category only (see, e.g., De Deyne et al., 2008). Studies employing free naming (Malt & Sloman, 2004) and name verification (Malt & Sloman, 2004; Ruts, Storms, & Hampton, 2004) also confirm that participants often award multiple labels to familiar artifact items. Hampton (1998) had participants endorse or discard items as category members. When he attempted to predict the proportion of times an item would be endorsed as a category member from that item's typicality rating, he found that underestimates for artifacts, but not for natural items, were often due to the item's membership in another category.

Even when artificially constructed stimuli, for which no a-priori-established membership convention can be said to exist, are presented for categorization, the domain difference is found. When asked to judge whether an item described as "halfway between" two categories (1) was probably one or the other, (2) could be called either one, or (3) couldn't be part of either category, participants in a study by Malt (1990) generally opted for the first alternative for natural items, while preferring the second alternative for many artifacts. Participants in Hampton et al. (2009) who were presented with chimerical creatures (e.g., a creature that has both crab and lobster features) for categorization, were more likely to place them in neither of the categories from which the features were inherited, than in both. Artifact items with hybrid features were more likely to be categorized in both categories than to be placed in neither.

These results support the idea that artifact groupings are best thought of as overlapping, while natural groupings are presumably more tightly clustered, a conclusion that is also arrived at in studies that are not restricted to membership of well-established, lexicalized categories. Ceulemans and Storms (2010) looked for latent structure in applicability matrices. The rows and columns of a domain's applicability matrix are made up of the domain's exemplars and characteristic features, respectively. Its entries indicate whether or not the features apply to the exemplars. The latent-class analysis of a matrix made up of animal exemplars and features yielded mutually exclusive classes: Each of the animal exemplars was awarded to just one of the classes. A matrix made up of a large selection of artifact exemplars and features, however, yielded classes that were intertwined. Several artifact exemplars were classified as belonging to multiple classes. Although the procedure that Ceulemans and Storms employed merely looked for dense regions in the applicability matrices, without any reference to category labels, it corroborated the discrete nature of natural groups and the overlapping nature of artifact groups that had been found in studies that did employ well-established category labels.

General notes on the experiments

Unlike natural items, artifacts do not seem uniquely segregated into groups (Keil et al., 2007; Malt & Sloman, 2007). The number of different groupings or contexts in which artifacts and natural items figure constitutes an uncontrolled variable in the paradigm that has traditionally been used to establish a domain difference in categorization. If context variety indeed shows a domain difference, as is suggested by the work reviewed above, it might explain earlier reports on domain differences in categorization. In addition, context variety has the potential to explain the categorization differences that have been found within domains. The first two experiments that we present investigated just that. Both entailed an analysis of the relationship between previously published categorization data (i.e., Estes, 2004) and newly gathered context variety judgments. The third experiment entailed an analysis of new categorization data; it differs from previous studies in that context variety did inform item selection.

All three experiments involved borderline category items, as they are most likely to receive graded membership responses (Diesendruck & Gelman, 1999; Estes, 2003, 2004; McCloskey & Glucksberg, 1978). Following Estes (2004), we defined artifact categories to be those that occur by human production or intention. Natural categories, on the other hand, were those that occur independently of human production or intention. Note, however, that the

Estes (2004) natural stimuli do not all adhere equally well to this definition. They include flowers, fruits, vegetables, dogs, and horses, which are all likely to be the result of intensive selective breeding, to the effect that they are now biological artifacts. The decision to adopt the Estes (2004) stimuli thus came with a somewhat broader range of natural stimuli than some other studies have employed. This might be part of the reason why Estes (2004) showed considerable categorization variability within the domain of natural items.

The reasons for using context variety judgments to evaluate the hypothesis are threefold. First of all, our context variety account of categorization differences assumes that participants actively search for other contexts of use for a stimulus than the one suggested by the target category. Context variety judgments presumably constitute the best approximation of this process. Second, the procedure doesn't require contexts to be restricted to lexicalized categories. Any context of use that is considered distinct from the one suggested by the target category may count toward the variety judgment. Third, unlike generation data, judgments on a Likert-type scale do not require subjective processing (e.g., judging whether the contexts generated by different participants are identical or not) and allow for greater data variability (for items for which one context is very dominant). Nevertheless, the context variety judgments we collected proved to be skewed. Therefore, we report nonparametric tests, although none of the conclusions are dependent on their use; use of the parametric counterparts of the Wilcoxon Mann–Whitney test, the Wilcoxon signed rank test, and Spearman's correlation coefficient supported the same conclusions.

Experiment 1

The purpose of the first experiment was twofold: to establish (1) whether context variety is greater for artifacts than for natural items and (2) whether differences in context variety parallel those in gradedness. To avoid stimulus-sampling issues, the materials for Experiment 1 were taken from an earlier study that established a categorization difference between natural and artifactual stimuli (i.e., Estes, 2003, 2004).

Method

A total of 16 undergraduates at City University London participated voluntarily in a context variety judgment task. The materials were borderline artifact and natural items that were previously used by Estes (2003, Exp. 1; 2004, Exp. 1). Five items were included for each of four artifact categories (furniture, tools, vehicles, weapons) and each of four natural categories (birds, fruits, trees, vegetables). All of the items that Estes chose had received an average membership rating between 3.01 and 5.00 in Barr and Caplan (1987). The scale in that particular study ranged from 1 for clear nonmembers to 7 for clear members. The items in the two domains were shown to match for typicality using both the original typicality norms provided in Barr and Caplan and newly gathered ratings by Estes (2004).

A measure of category gradedness was available from Estes (2004). Participants were asked to respond to each pairing of an item with its target category using a membership scale that ranged from 0 (*not at all a member*) to 10 (*completely a member*). The proportion of non-endpoint responses (i.e., any rating from 1 to 9) served as the measure of category gradedness.

The instructions for the context variety judgment task were taken from Galbraith and Underwood (1973), who used the variable to explain a divergence between the perceived and objective frequencies of abstract and concrete words. Galbraith and Underwood carefully tested the instructions to ensure that participants would not be inclined to judge word frequency. We slightly adapted their original instructions so that they would refer to language use in general instead of being restricted to written discourse. The instructions read as follows:

> Words differ widely in the variety of contexts with which they are used. The word *skillet* (the participants were told) has a very limited number of contexts, since the word would probably nearly always have reference to a kitchen. The word *water*, on the other hand, may be used in several different contexts, such as references to water in the well, water bed, mineral water, and so on. We would like you to read through the words below and rate each word on the 9-point scale according to how varied the contexts are in which the word is used.

The 9-point rating scale that participants had at their disposal featured the anchor words *skillet* (1) and *water* (9). Note that, if anything, these instructions ran counter to the direction of the effect we hoped to find.

Participants were further discouraged from making guesses for words that they did not know, but were asked to circle the words instead. Half of the participants saw the items in alphabetical order, and the other half saw them in reverse alphabetical order.

Results

Figure 1 depicts the gradedness measure that was taken from Estes (2004) and the newly gathered context variety judgments.

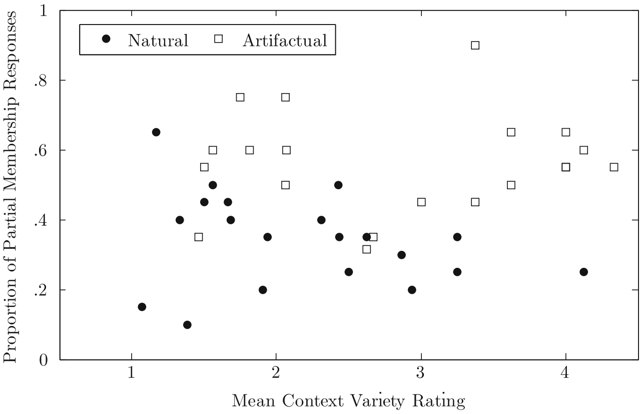

Fig. 1 Gradedness as a function of context variety, Experiment 1

Context variety The averaged context variety ratings were submitted to two analyses; one with participants, and another with items treated as random. The Wilcoxon Mann–Whitney test was used in the items analysis, and the Wilcoxon signed rank test was employed in the participants analysis. Both the former (Z = 2.12, N = 20, p < .05) and the latter (S = 47, N = 16, p < .01) indicated context variety to be judged higher for artifacts than for natural items. Inspection of Fig. 1 reveals that the effect is due to about half of the artifact items (i.e., white squares) extending beyond the cloud of natural items (i.e., black circles) along the horizontal context variety axis.

Context variety and gradedness Context variety and gradedness were not positively correlated either within the artifact domain (ρ = .06, N = 20, p = .39) or within the natural domain (ρ = −.29, N = 20, p = .89). Nor were they positively correlated across all items (ρ = .13, N = 40, p = .22).

Discussion

Experiment 1 provides evidence for the predicted domain difference in context variety, in that the artifacts were judged to appear in a greater variety of contexts than the natural items. However, the results do not support the context variety account of categorization differences. Across all stimuli, the relation between context variety and gradedness was positive but did not reach significance. There was no significant positive relation within the artifact or the natural domain, either.

Experiment 2

Experiment 1 established a domain difference in context variety: Artifacts were judged to feature in a greater variety of contexts than natural items. The difference did not coincide with the domain difference in categorization, however. Nor could context variety explain the categorization differences within a single domain. The purpose of the second experiment was to establish whether these results would generalize to another set of items. Another experiment by Estes (2004) offered materials and data that would allow for such an investigation. It too employed a selection of borderline artifact and natural items for which a domain difference in categorization had been established. With respect to the materials of the first experiment, it has the added benefit that (1) it includes a greater selection of items, (2) the borderline items were sampled in a more heterogeneous manner, and (3) the employed category membership judgment task more explicitly addressed the notion of graded membership.

Method

A total of 39 undergraduates at City University London participated voluntarily in a context variety judgment task. The materials were borderline artifact and natural items that had previously been used by Estes (2003, Exp. 2; 2004, Exp. 2). These included the items that Kalish (1995) intuitively deemed borderline, another set of items from Barr and Caplan (1987), and a further set of items taken from McCloskey and Glucksberg (1978). The items that were taken from Barr and Caplan again received an average membership rating between 3.01 and 5.00 on the 7-point scale they used. These items were different from the ones we used in Experiment 1. The items from McCloskey and Glucksberg were selected as those having elicited the most disagreement in the binary (yes/no) categorization task the researchers had conducted. The selected items were all associated with percentages of nonmodal responses between 30% and 50%. This led to the inclusion of 39 artifact and 39 natural borderline items. They were paired with their corresponding target categories (clothing, furniture, kitchen utensils, ships, toys, and weapons in the artifact domain; animals, dogs, fish, flowers, horses, insects, and mammals in the natural domain) to make up the final set of materials. Note that, unlike the stimuli in Experiment 1, the artifact and natural items in Experiment 2 were never matched for typicality.

The measure of category gradedness available from Estes (2004) was a proportion of partial-member choices. Participants were asked to select one of three alternatives for each pairing of an item with its target category: nonmember, partial member, or full member. Partial membership was taken to mean that the item belonged in the target category, but not to the same extent as some other items. This procedure to elicit membership judgments is believed to be superior to that employed in the first experiment, in that it explicitly addresses the graded-membership notion (Estes, 2003).

The instructions and procedure for the context variety judgment task were identical to those used in Experiment 1.

Results

Figure 2 depicts the measures of gradedness and context variety.

Context variety The averaged context variety ratings were again submitted to two analyses. The Wilcoxon Mann–Whitney test was used for the items analysis, and the Wilcoxon signed rank test for the participants analysis. Artifacts appeared in a greater variety of contexts, both according to the items analysis ($Z = 3.01$, $N = 39$, $p < .01$) and the participants analysis ($S = 354.5$, $N = 39$, $p < .0001$). The effect is apparent in Fig. 2, where a considerable number of artifact items (i.e., white squares) are located to the right of the natural items (i.e., black circles) along the horizontal context variety axis. As compared to the artifact items, the range of the natural items with respect to context variety is clearly restricted.

Context variety and gradedness Context variety and gradedness were not positively correlated within the artifact domain ($\rho = -.09$, $N = 39$, $p = .71$) or within the natural domain ($\rho = -.06$, $N = 39$, $p = .65$). Nor were they positively correlated across all items ($\rho = .15$, $N = 78$, $p = .10$).

Discussion

The results of Experiment 2 corroborate those of Experiment 1. As in Experiment 1, the artifacts were found to score higher on context variety than the natural items, thus indicating a context variety domain difference. The relation between context variety and gradedness, however, did not reach significance across items, within artifacts, or within natural items in this experiment either. Gradedness and context variety do not display the positive relationship that the context variety account of categorization differences posits.

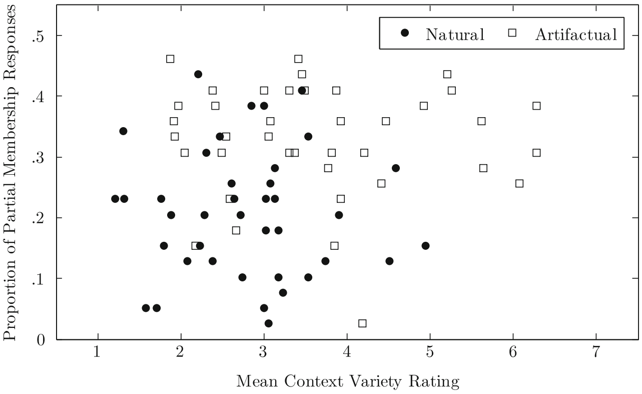

Fig. 2 Gradedness as a function of context variety, Experiment 2

Experiment 3

The first two experiments showed little evidence for a relationship between context variety and gradedness. Although both experiments demonstrated domain differences in context variety as well as gradedness, with artifacts scoring higher than natural items on both measures, no consistent relationship between the two variables was demonstrated. One might wonder whether this rules out any influence of context variety on gradedness. To avoid stimulus-sampling issues, the materials in Experiments 1 and 2 were taken from previous studies in which the focus was on establishing domain differences in gradedness. The stimulus selection procedure was thus uninformed with respect to context variety. Experiments 1 and 2 suggest that the context variety population mean of the borderline artifacts is higher than that of the borderline natural items. This would explain the significant domain difference in context variety following two stimulus selection procedures that are not in any way guided by the variable's distributional properties.

However, this does not preclude the existence of individual items that violate the population-level tendencies—natural items that are as high on context variety as most artifacts, and artifacts with the low context variety of most natural items. The bottom left corner of Fig. 1, for instance, shows an artifact (gondola) that received an average context variety score that resembles that of the lowest-scoring natural items. The item also scores relatively low on gradedness. This raises the question of whether an influence of context variety on gradedness can be established through an informed selection of stimuli (i.e., through the reversal of the domain difference that was established in the first two experiments). Under such "irregular" conditions, participants may be more prone to demonstrate an influence of context variety. The aim of the third experiment was to assess this hypothesis.

In order to identify items that might violate the regular distribution of context variety across artifacts and natural kinds, we turned to text corpora studies. Many of these corpora, such as the Touchstone Applied Science Associates corpus (TASA; Landauer, Foltz, & Laham 1998), are made up of semantically coherent text documents. It has been argued that the number of such documents a word appears in is a proxy of the variety of contexts it is used in (Adelman, Brown, & Quesada 2006; Steyvers & Malmberg, 2003). These document counts have been shown to relate to the words' number of meanings (Adelman et al., 2006). Experiment 3 proceeded in several steps. We selected a number of borderline artifacts and natural items that were matched for typicality, but in terms of context variety were thought to display a pattern opposite to that demonstrated in the first two experiments: According to the TASA document counts, the selected natural items appeared in more contexts than did the artifacts. To ensure that the context variety pattern was actually reversed for these items, we then had a group of undergraduates perform a context variety judgment task. Finally, another group of undergraduates was asked to categorize each of these items as a nonmember, a partial member, or a full member of a suggested target category.

Method

A total of 18 undergraduates at the University of Wisconsin–Madison participated in a context variety judgment task for partial course credit. Then, 32 different undergraduates at the same university participated in a categorization task for partial course credit.

Following the inclusion criteria that informed the stimulus selections in Experiments 1 and 2, stimuli were sampled from Barr and Caplan (1987) and from McCloskey and Glucksberg (1978). All of the stimuli could be considered borderline, in that they either received a mean membership rating between 3.01 and 5.00 in the Barr and Caplan norms or had a mean proportion of nonmodal responses of .25−.50 in the McCloskey and Glucksberg norms. Note that the latter constitutes a small deviation from the original inclusion criterion, which only allowed for stimuli with a mean nonmodal response proportion of .30−.50. This deviation was required in order to find a decent-sized set of borderline stimuli (20 artifacts and 20 natural items) that, while being matched with regard to the typicality judgments of Barr and Caplan and McCloskey and Glucksberg, would be expected to show a reversal of the context variety domain difference found in the first two experiments. The reversal was apparent in the higher log-transformed TASA document count for the natural items than for the artifacts.

The Appendix lists the selected items, along with their respective target categories. The artifact domain encompasses the categories of carpenter's tools, clothing, furniture, kitchen utensils, ships, vehicles, and weapons. The natural domain encompasses the categories of animals, fish, fruits, insects, mammals, trees, and vegetables.

One group of undergraduates was then asked to judge the variety of contexts in which the selected natural and artifactual items appeared. The instructions for the context variety judgment task were identical to those used in Experiments 1 and 2. The 40 items were randomly ordered, and the item presentation either followed this random ordering or its reverse. The aim of the task was to validate the stimulus selection procedure involving the TASA norms and to obtain a measure of context variety that would be as similar as possible to the one that would (potentially) inform participants' categorization decisions.

The stimuli were also presented for categorization to another group of undergraduates. The items were presented in a random order or in the reverse order. Participants could choose to indicate an item as a nonmember, a partial member, or a full member of its suggested target category. In accordance with the original instructions by Estes (2003, 2004), participants were told that a choice of partial membership meant that the item belonged in the target category, but only to a degree, not fully. The proportion of partial-membership choices served as the measure of category gradedness.

Results

Gradedness and context variety are depicted in Fig. 3.

Context variety When all 40 stimuli were included, both the items analysis ($Z = -1.65$, $N = 20$, $p = .05$) and the participants analysis ($S = -41.5$, $N = 18$, $p = .05$) did not provide clear indications of the reliability of the context variety domain difference. Therefore, we decided to remove the two items (one in each domain) that most clearly violated the difference that was aimed for. When the natural item with the lowest context variety score (*octopus*) and the artifactual item with the highest context variety score (*satellite*) were removed, both the Wilcoxon Mann–Whitney test ($Z = -2.32$, $N = 19$, $p < .05$) and the Wilcoxon signed rank test ($S = -53.5$, $N = 18$, $p < .05$) indicated that the stimulus selection procedure was successful. Contrary to the pattern found in Experiments 1 and 2, the natural domain was now the one displaying the higher mean context variety.

Typicality With *octopus* and *satellite* removed from the stimulus set, the natural and artifactual domains were still matched for typicality. The typicality ratings provided in Barr and Caplan (1987, 7-point rating scale) and

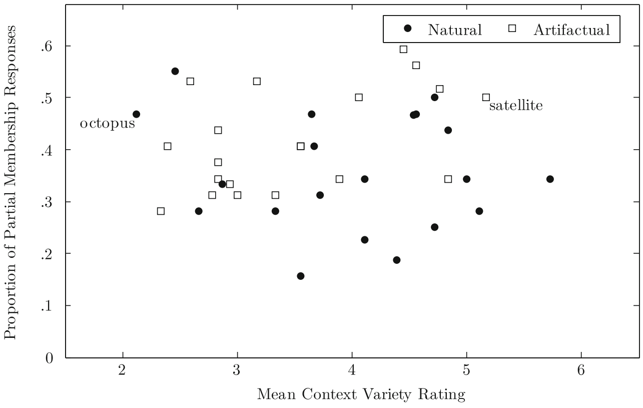

Fig. 3 Gradedness as a function of context variety, Experiment 3

McCloskey and Glucksberg (1978, 10-point rating scale) were brought onto a common scale by first subtracting 1 and then dividing by the maximum value of the respective rating scale minus 1. The resulting values were subjected to the Wilcoxon Mann–Whitney test ($Z = 0.99$, $N = 19$, $p = .32$). A participants analysis could not be conducted, because the necessary data were not available in the accompanying manuscripts.

Gradedness Despite the reversal of the context variety domain difference, the items analysis ($Z = 1.74$, $N = 19$, $p < .05$) still showed that participants were more inclined to categorize the artifacts in a graded manner than the natural items. The proportions of partial-membership responses did not differ from one domain to the other according to the participants analysis ($S = -43$, $N = 32$, $p = .38$).

Context variety and gradedness Context variety and gradedness were not positively correlated across all items ($\rho = .07$, $N = 38$, $p = .33$). Nor were they positively correlated within the natural domain ($\rho = .07$, $N = 19$, $p = .39$). Within the artifact domain, their relationship was somewhat more pronounced ($\rho = .39$, $N = 19$, $p = .05$).

Discussion

The results from the third experiment yielded little evidence for the context variety account of categorization differences. Contrary to the account's predictions, a selection of artifacts that was lower in context variety than a selection of natural items still did not elicit fewer graded membership judgments. Even under conditions that were specifically created to bring about an effect of context variety on categorization, participant performance was more in accord with the existing representational accounts of categorization differences. Indeed, the domain difference in categorization that the representational accounts posit reached significance in the items analysis reported above, and even when the stimulus set was restricted to the 15 or 10 artifacts and natural items that best displayed the reversal of the context variety domain difference found in Experiments 1 and 2, the average gradedness score was higher for artifacts than for natural items. In none of these additional analyses did we find evidence for a positive relationship between context variety and gradedness, within or across domains.

General discussion

We conducted three studies to test the hypothesis that the more contexts an item appears in, the more graded the categorization responses it will elicit. In all three experiments, we failed to find a positive relationship between context variety and gradedness. The context variety hypothesis was put forward to explain the results of categorization studies that have presented categorization differences within and (most notably) between the domains of artifacts and natural kinds. Clearly, the present results do not support the context variety account as a viable alternative for existing representational accounts of these categorization differences. Limited though the predictive power of these representational accounts may be, they can at least nominally account for the intra- and interdomain differences in gradedness, while context variety cannot.

One might object that the null result is due to a confounding of context variety and familiarity. Earlier work has established a strong relationship between judgments of context variety and judgments of familiarity (Galbraith & Underwood, 1973). Data from a small pilot study in which 9 participants judged how familiar they were with the items from Experiment 1 suggested that this relationship also holds for the present stimuli. The Spearman correlation between the mean context variety and the mean familiarity judgments was established at .59 ($N = 40$, $p < .0001$). Due to the correlational nature of this result, one cannot ascertain whether differences in familiarity are responsible for the corresponding differences in context variety or whether the context variety differences are responsible for the familiarity differences. With respect to the evaluation of the context variety account of gradedness, an influence of familiarity on judgments of context variety should not be considered a cause for concern. As was stated in the introduction, a similar process was thought to inform both graded categorization and judgments of context variety. A deliberate consideration of the different contexts in which an item can occur was thought to precede categorization. If familiarity with the item is a major determinant of the variety of contexts it is judged to appear in, familiarity would presumably assert a similar influence on the search for contexts prior to a categorization decision. For this reason, we also chose to evaluate the relationship between context variety and gradedness using participants' judgments instead of the corpus-derived measure that was introduced in Experiment 3.

A domain difference in context variety

The present results are the first to explicitly establish a domain difference in context variety, with artifact items judged to appear in a greater variety of contexts than natural items. It was shown to be a marked difference in two sets of stimuli that were compiled without regard of this variable. Moreover, when we required a set of borderline artifacts that was lower in context variety than a set of borderline natural stimuli, it proved quite difficult to find exceptions to the domain difference. The original inclusion criteria needed to be loosened to find a decent-sized sample of exceptions. This suggests that the domain difference is a pervasive one. In order to strengthen this finding, we investigated four extensive stimulus sets to see whether the included artifacts and natural items would also demonstrate the context variety difference. Because of the size of the stimulus sets, we decided to employ a corpus-derived measure instead of context variety judgments. We chose to use a different measure than the one that informed the stimulus selection in Experiment 3. The reason for this is that just as judgments of context variety and familiarity show a strong relationship, counts of the number of documents a word appears in and word frequency may be intimately related: More frequent words must, by definition, occur in more documents (Hoffman, Rogers, & Lambon Ralph, 2011). For the purpose of comparing artifacts and natural items with respect to dependent variables other than categorization, it might be of importance to disentangle these variables.

Hoffman, Rogers, and Lambon Ralph (2011) devised a means of obtaining a context variety measure from corpus data that is, at least in principle, independent of word frequency (it might still turn out that low-frequency words are found to be low in context variety, and high-frequency words to be high in context variety). At the root of the proposed method is the key principle of latent semantic analysis—namely, that the context in which a word is found (e.g., a sample of texts on a particular topic) carries information about its meaning (Landauer & Dumais, 1997). This intuition is generally applied to determine the semantic similarity between two or more words: They are considered similar in meaning to the extent that their contexts resemble one another. By applying the same rationale to the contexts of a single word, one can derive a measure of context variety. If the average similarity of the various contexts in which the word appears is high, the word is considered to be low in context variety. Likewise, the word is considered to be high in context variety when the average similarity of its contexts is low. The average similarity among associated contexts is in principle completely independent of a word's frequency.

Hoffman et al. (2011) applied this procedure to the written text portion of the British National Corpus (British National Corpus Consortium, 2007), log transformed the average context similarity measures, and reversed the sign to yield a measure of context variety. They have made context variety norms available for 12,618 English words. Among these are 138 artifacts and 156 natural items from the three reports that informed the choice of stimuli for the experiments in the present study (Barr & Caplan, 1987; Kalish, 1995; McCloskey & Glucksberg, 1978). According to the Hoffman measure of context variety, these artifacts appear in a greater variety of contexts than the natural items ($Z = 2.18$, $p < .05$). The same conclusion was reached for 105 artifacts and 62 natural items from Diesendruck and Gelman (1999). Context variety is higher for artifacts than for natural kinds in this sample of items from one of the first studies to have established a domain difference in categorization ($Z = -2.17$, $p < .05$). From Hampton and Gardiner (1983), 108 artifacts and 131 natural items are available. The artifacts again score higher on context variety than the natural items do ($Z = 1.79$, $p < .05$). Finally, from the Cree and McRae (2003) norms, 256 artifacts and 137 natural items are available. Once again, the artifactual items attain higher scores on the context variety measure than do the natural items ($Z = -1.65$, $p < .05$). These results further establish the domain difference in context variety that we obtained using participant judgments. In four sets of commonly employed stimuli, the artifact items are found to appear in a greater variety of contexts than the natural items using a measure that is not confounded with subjective or objective frequency. Although an investigation into the origins of this domain difference is beyond the scope of this study, it is worth pointing out that instances of artifact categories in recent years have undergone (and still continue to undergo) considerable changes due to all sorts of technological advances, while the role that instances of natural categories play in our lives has been much more consistent (see Malt, 2010, for a more elaborate discussion of these matters).

Implications

We see at least three possible manners in which this finding may influence future work. This particular domain of study is believed to be subject to many influences that, depending on how studies are set up, may or may not present themselves (Malt & Sloman, 2007). Many have therefore argued for more rigorous control of the stimulus materials employed (Estes, 2003; Kalish, 1995, 2002). We believe context variety to be one of those influences that one would want to control for, either statistically or experimentally, in order to eliminate it as a confounding factor. Alternatively, one could altogether abandon the custom of presenting items and target categories in relative isolation, without any reference to a meaningful (discourse) context. Instead, one could, for instance, embed the items in meaningful sentence contexts that provide cues to their identity. This has been a common practice in the studies by Schwanenflugel and colleagues on context availability (e.g., Schwanenflugel & Shoben, 1983; Schwanenflugel & Stowe, 1989). Their work conveys the idea that the greater the amount of information associated with a particular item, the more difficult it is to retrieve any particular piece of that information. Providing an appropriate context might then make the category-ness of an item more salient. Likewise, one might want to abandon the use of underspecified items. We use the word *knife* in a variety of circumstances, sometimes indicating it to be a piece of tableware, while on other occasions we refer to it as a weapon or a tool. The same is not true of individual knives, which are presumably all associated with one or two specific contexts. For instance, many individual weapon knives would never be considered tableware knives. It is to be expected that items that appear in a variety of contexts, like most artifacts, would be most subject to such methodological changes. If their identity is indeed context dependent (as the work reviewed in this report leads us to believe), they might present with significantly less gradedness—similar to that elicited by natural items with conventional (context-free) identities—in a clearly specified categorization context or when a presentation modality (e.g., photographs) is used that doesn't leave the item underspecified. The observation that, relative to a particular context or presentation mode, categorization is all-or-none, *regardless* of the type of item involved, would pose a serious challenge to those accounts that propose that representational differences account for differences in categorization (Braisby, Franks, & Hampton 1996; Franks, 1995).

Appendix

Table 1 Experiment 3 stimuli

Artifact		Natural	
Category	Item	Category	Item
carpenter's tools	calculator	animals	bacterium
carpenter's tools	varnish	animals	yeast
clothing	corduroy	fish	clam
clothing	handbag	fish	crab
clothing	necklace	fish	octopus
clothing	wig	fish	shrimp
furniture	candlestick	fruits	coconut
furniture	mantel	fruits	olive
kitchen utensils	broom	fruits	pumpkin
kitchen utensils	dishwasher	fruits	tomato
kitchen utensils	mop	insects	spider
ships	canoe	insects	worm
ships	kayak	mammals	goose
ships	raft	trees	hemlock
ships	spacecraft	trees	lilac
vehicles	escalator	trees	sage
vehicles	parachute	vegetables	gourd
vehicles	stretcher	vegetables	peanut
vehicles	surfboard	vegetables	rice
weapons	satellite	vegetables	sugarcane

References

Adelman, J. S., Brown, G. D. A., & Quesada, J. F. (2006). Contextual diversity, not word frequency, determines word naming and lexical decision times. *Psychological Science, 17,* 814–823.

Barr, R. A., & Caplan, L. J. (1987). Category representations and their implications for category structure. *Memory & Cognition, 15,* 397–418.

Bear, G. (1974). Implicit alternatives to a stimulus, difficulty of encoding, and schema-plus-correction representation. *Memory & Cognition, 2,* 360–366.

Block, N. (1998). Semantics, conceptual role. In E. Craig (Ed.), *Routledge encyclopedia of philosophy, vol. 8* (pp. 652–657). London: Routledge.

Braisby, N. R., Franks, B., & Hampton, J. A. (1996). Essentialism, word use, and concepts. *Cognition, 59,* 247–274.

British National Corpus Consortium. (2007). *British national corpus version 3* (BNC XMLth ed.). Oxford: Oxford University Computing Services.

Ceulemans, E., & Storms, G. (2010). Detecting intra- and inter-categorical structure in semantic concepts using HICLAS. *Acta Psychologica, 133,* 296–304. doi:10.1016/j.actpsy.2009.11.011.

Cree, G., & McRae, K. (2003). Analyzing the factors underlying the structure and computation of the meaning of chipmunk, cherry, chisel, cheese, and cello (and many other such concrete nouns). *Journal of Experimental Psychology. General, 132,* 163–201.

De Deyne, S., Verheyen, S., Ameel, E., Vanpaemel, W., Dry, M. J., Voorspoels, W., et al. (2008). Exemplar by feature applicability matrices and other Dutch normative data for semantic concepts. *Behavior Research Methods, 40,* 1030–1048.

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

Estes, Z. (2003). Domain differences in the structure of artifactual 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.

Franks, B. (1995). Sense generation: A "quasi-classical" approach to concepts and concept combination. *Cognitive Science, 19,* 441–505.

Galbraith, R. C., & Underwood, B. J. (1973). Perceived frequency of concrete and abstract words. *Memory & Cognition, 1,* 56–60.

Garner, W. R. (1966). To perceive is to know. *The American Psychologist, 21,* 11–19.

Gelman, S. A., & Hirschfeld, L. A. (1999). How biological is essentialism? In D. Medin & S. Atran (Eds.), *Folkbiology* (pp. 403–446). Cambridge, MA: MIT Press.

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. (2009). 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* (pp. 80–99). Oxford: Oxford University Press.

Hampton, J. A., & Gardiner, M. M. (1983). Measures of internal category structure: A correlational analysis of normative data. *British Journal of Psychology, 74,* 491–516.

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

Hampton, J. A., Estes, Z., & Simmons, S. (2007). Metamorphosis: Essence, appearance, and behavior in the categorization of natural kinds. *Memory & Cognition, 35,* 1785–1800.

Hampton, J. A., Storms, G., Simmons, C. L., & Heussen, D. (2009). Feature integration in natural language concepts. *Memory & Cognition, 37,* 1150–1163.

Hoffman, P., Rogers, T. T., Lambon Ralph, M. A. (2011). Semantic diversity accounts for the "missing" word frequency effect in stroke aphasia: Insights using a novel method to quantify contextual variability in meaning. *Journal of Cognitive Neuroscience*

Kahneman, D., & Miller, D. T. (1986). Norm theory: Comparing reality to its alternatives. *Psychological Review, 93,* 136–153.

Kalish, C. W. (1995). Essentialism and graded membership in animal and artifact categories. *Memory & Cognition, 23,* 335–353.

Kalish, C. W. (2002). Essentialist to some degree: Beliefs about the structure of natural kind categories. *Memory & Cognition, 30,* 340–352.

Keil, F. C., Greif, M. L., & Kerner, R. S. (2007). A world apart: How concepts of the constructed world are different in representation and in development. In E. Margolis & S. Laurence (Eds.), *Creations of the mind: Theories of artifacts and their representation* (pp. 232–245). New York: Oxford University Press.

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review, 104,* 211–240. doi:10.1037/0033-295X.104.2.211.

Landauer, T. K., Foltz, P. W., & Laham, D. (1998). Introduction to latent semantic analysis. *Discourse Processes, 25,* 259–284. doi:10.1080/01638539809545028.

Malt, B. C. (1990). Features and beliefs in the mental representation of categories. *Journal of Memory and Language, 29,* 289–315.

Malt, B. C. (2010). Naming artifacts: Patterns and processes. In B. H. Ross (Ed.), *The psychology of learning and motivation: Advances in research and theory, vol. 52* (pp. 1–38). San Diego: Academic Press.

Malt, B. C., & Sloman, S. A. (2004). Conversation and convention: Enduring influences on name choice for common objects. *Memory & Cognition, 32,* 1346–1354.

Malt, B. C., & Sloman, S. A. (2007). Artifact categorization: The good, the bad, and the ugly. In E. Margolis & S. Laurence (Eds.), *Creations of the mind: Theories of artifacts and their representation* (pp. 85–123). New York: Oxford University Press.

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

Rhodes, M., & Gelman, S. A. (2009). Five-year-olds' beliefs about the discreteness of category boundaries for animals and artifacts. *Psychonomic Bulletin & Review, 16,* 920–924.

Ruts, W., Storms, G., & Hampton, J. A. (2004). Linear separability in superordinate natural language concepts. *Memory & Cognition, 32,* 83–95.

Schwanenflugel, P. J., & Shoben, E. J. (1983). Differential context effects in the comprehension of abstract and concrete verbal materials. *Journal of Experimental Psychology. Learning, Memory, and Cognition, 9,* 82–102.

Schwanenflugel, P. J., & Stowe, R. W. (1989). Context availability and the processing of abstract and concrete words in sentences. *Reading Research Quarterly, 24,* 114–126.

Steyvers, M., & Malmberg, K. J. (2003). The effect of normative context availability on recognition memory. *Journal of Experimental Psychology. Learning, Memory, and Cognition, 29,* 760–766.

Tversky, A., & Koehler, D. J. (1994). Support theory: A nonextensional representation of subjective probability. *Psychological Review, 101,* 547–567.

Windschitl, P. D., & Wells, G. L. (1998). The alternative-outcomes effect. *Journal of Personality and Social Psychology, 75,* 1411–1423.

Wittgenstein, L. (1953). *Philosophical investigations (Trans. G. E. M. Anscombe)*. Oxford: Blackwell.