Acta Psychologica 172 (2017) 26–40

Contents lists available at ScienceDirect

Acta Psychologica

journal homepage: www.elsevier.com/locate/actpsy

Free classification of large sets of everyday objects is more thematic than taxonomic

Rebecca Lawson [a,*], Franklin Chang [a], Andy J. Wills [b]

[a] Department of Experimental Psychology, University of Liverpool, UK
[b] School of Psychology, Plymouth University, UK

ARTICLE INFO

Article history:
Received 16 March 2016
Received in revised form 25 August 2016
Accepted 1 November 2016
Available online 15 November 2016

Keywords:
Semantic knowledge
Unsupervised categorization
Free-sorting
Concept

ABSTRACT

Traditionally it has been thought that the overall organisation of categories in the brain is taxonomic. To examine this assumption, we had adults sort 140–150 diverse, familiar objects from different basic-level categories. Almost all the participants (80/81) sorted the objects more thematically than taxonomically. Sorting was only weakly modulated by taxonomic priming, and people still produced many thematically structured clusters when explicitly instructed to sort taxonomically. The first clusters that people produced were rated as having equal taxonomic and thematic structure. However, later clusters were rated as being increasingly thematically organised. A minority of items were consistently clustered taxonomically, but the overall dominance of thematically structured clusters suggests that people know more thematic than taxonomic relations among everyday objects. A final study showed that the semantic relations used to sort a given item in the initial studies predicted the proportion of thematic to taxonomic word associates generated to that item. However, unlike the results of the sorting task, most of these single word associates were related taxonomically. This latter difference between the results of large-scale, free sorting tasks versus single word association tasks suggests that thematic relations may be more numerous, but weaker, than taxonomic associations in our stored conceptual network. Novel statistical and numerical methods for objectively measuring sorting consistency were developed during the course of this investigation, and have been made publicly available.

Organising our knowledge of the world into useful categories is one of the brain's most basic functions. Categories are collections or classes of objects or entities that are similar or related in some meaningful way. Two types of relation, taxonomic and thematic, have been widely proposed to provide structure to our stored, semantic knowledge about categories of concrete objects such as trees and hammers (Murphy, 2002). Taxonomic relations group together the same kinds of objects based on perceptual and functional similarities (De Deyne, Verheyen, & Storms, 2016; Rosch, Mervis, Gray, Johnson, & Boyes-Braem, 1976). As an example, most members of the category *fruit* have similar shapes, sizes, smells and tastes and they are grown and used in similar ways. In contrast, thematic relations group together objects that need not be either perceptually or functionally similar to each other. Instead they normally have complementary roles in events and co-occur in common situations, locations and/or times (De Deyne et al., 2016; Lin & Murphy, 2001). For example, the thematic category *mail* might include an envelope, a post-box, a stamp, a post-man and a parcel.

Adults can easily categorize objects both taxonomically and thematically and they can vary this behaviour in response to task demands (Estes, Golonka, & Jones, 2011; Wisniewski & Bassok, 1999; Shafto, Kemp, Mansinghka, & Tenenbaum, 2011). Thus, at a fine scale of analysis, both types of semantic relation are readily available. However, at a coarse scale our long-term, semantic knowledge is often thought to be organised taxonomically because taxonomic relations reflect the deep, causal structure of our world and so can support useful inductions. In addition, much of our formal education may encourage us to categorize taxonomically (Estes et al., 2011). Thematic relations arise from storing our episodic experiences of the world and, specifically, co-occurrences in time and space. It has been argued that such relations may be less useful as a basis for induction. There is, however, little empirical evidence to evaluate the claim that, overall, our stored semantic knowledge is mainly taxonomic rather than thematic. The present study investigated this claim using a large-scale, open-ended free-sorting task where adults were asked to cluster sets of objects which go naturally together. We investigated whether these clusters were consistent across different people (rather than being idiosyncratic) and, if so, what type of semantic relations (taxonomic versus thematic) was principally used to structure them.

Several studies have suggested that adults prefer to categorize taxonomically (e.g., Olver & Homsby, 1966; Ross & Murphy, 1999; Smiley &

* Corresponding author at: Department of Experimental Psychology, University of Liverpool, Eleanor Rathbone Building, Bedford Street South, Liverpool L69 7ZA, UK.
E-mail address: rlawson@liverpool.ac.uk (R. Lawson).

http://dx.doi.org/10.1016/j.actpsy.2016.11.001

Brown, 1979; Tare & Gelman, 2010). However, Murphy (2001, 2002) noted that many previous free-sorting studies presented small numbers of items that could easily be sorted into salient taxonomic clusters (with no leftover items) but that could not so readily be organised into thematic clusters. Murphy suggested that the results of such studies may have led to an overestimation of adult's use of taxonomic relations to categorize. Consistent with this claim, Lin and Murphy (2001); see also Koriat & Melkman, 1981; Saalbach & Imai, 2007) found that when salient thematic relations were provided, many people consistently matched thematically rather than taxonomically. To further investigate this issue, Murphy (2001) used a small-scale, free-sorting task in which a set of nine pictures could be divided into either three taxonomic clusters (vehicles, professions and locations) or three thematic clusters (with themes of travel by aeroplane, boat or car), with three items per cluster in both cases. Participants were told to group the pictures in the way that seemed "best and most natural". Murphy found that most people sorted the stimuli thematically, suggesting that if both thematic and taxonomic relations are readily available then adults do not show a strong inclination to sort taxonomically. Murphy also provided evidence of the effect of stimulus selection in a follow-up experiment in which the three location objects were replaced by three animals. This meant that there were still three salient and equal-sized taxonomic clusters but that it was difficult to sort all of the items into thematically organised clusters. Most people now sorted taxonomically, demonstrating that the semantic relations used in free-sorting are both flexible and sensitive to the stimuli provided. Overall, Murphy's data shows that the results of many previous free-sorting studies, which had originally been interpreted as revealing that adults mainly organise their semantic knowledge taxonomically, could instead have arisen from experimenters selecting items that were more strongly related taxonomically than thematically.

1. A large-scale, free-sorting task

Research to date thus suggests that, first, people can flexibly use both taxonomic and thematic relations when grouping items, and that, second, many items can be readily grouped either taxonomically or thematically (Lin & Murphy, 2001; Murphy, 2001; Nguyen & Murphy, 2003; Smiley & Brown, 1979). This research is based on matching and small scale free-sorting studies which investigated categorization preferences at a fine-grained, detailed scale. However, this research is not informative about the overall nature of our stored semantic knowledge since experimenters usually selected stimuli to be either clearly taxonomically or clearly thematically related. Careful stimulus selection is necessary for such studies as the aim is to directly compare people's *choice* to respond based on taxonomic versus thematic relations of matching strength. A different and more coarse-grained method must be used to assess the relative number of taxonomic versus thematic relations *available* across a broad range of everyday items in order to infer the overall nature of our stored, semantic knowledge.

The present study investigated whether more thematic than taxonomic relations were available across basic level categories (such as apple, bowl and scissors; Lawson & Jolicoeur, 2003; Rosch et al., 1976) using a relatively unconstrained, large-scale, free-sorting task. We analysed how people clustered sets of 140–150 diverse, concrete objects. The large sets of items discouraged participants from assuming that there was a single, pre-defined, "correct" solution that the experimenter expected them to produce. This more exploratory, open-ended task provided considerable freedom for participants to choose the size and number of clusters to create and the semantic relations used to structure them. We assessed the consistency of clustering and the type of semantic relation used to cluster in order to provide evidence about whether the relations stored in our semantic knowledge are mostly idiosyncratic, taxonomic or thematic.

Several studies have examined the sorting of relatively large sets of objects but, unlike the present study, these have focussed on restricted domains of knowledge. For example, Medin et al. (2006); see also Medin, Lynch, Coley, & Atran, 1997) compared how two groups of experts (who lived in the same area but came from different cultures) clustered together 44 fish species. Sorting was similar overall but the Native American group used more ecological information, such as shared habitat, whereas the European American group was more likely to sort using biological class. Although knowledge was similar across the two cultures it was organised in different ways - more thematically than taxonomically for the Native Americans compared to the European Americans. Furthermore, Medin, Ross, Atran, Burnett, and Blok (2002) found that non-expert Native Americans and European Americans provided fewer taxonomic, and more goal-related, reasons for sorting when tested on the same task. Thus both culture and levels of expertise influenced the nature of semantic knowledge for this domain of fish species. Similarly, Lopez, Atran, Coley, Medin, and Smith (1997) found that USA undergraduates and indigenous Guatemalans both sorted around 40 mammals into categories similar to a standard scientific taxonomy. However, the USA students relied more on size to justify their sorting whereas the Guatemalans used a broader range of information and depended more on ecological relations. This, again, suggests that cultural differences can influence sorting. Note, though, that these studies only tested sorting for narrow domains of biological organisms which Western undergraduates are taught to think about taxonomically (Estes et al., 2011). Thus the responses to these single-domain sets of items may not generalise to the large, diverse sets of items that we used.

Recently, computational models have been shown to be able to extract and organise the types of semantic knowledge available to people using both thematic and taxonomic relations. For instance, natural language processing techniques such as latent semantic analysis (LSA) analyse the co-occurrence of words in text and can detect thematic relations (Landauer, McNamara, Dennis, & Kintsch, 2013). In contrast, deep learning approaches used in image recognition software such as GoogLe Net use visual similarity (features such as colour, size and shape) and can detect taxonomic relations (Szegedy et al., 2014). Recent modelling has suggested that a single mechanism can extract both types of relations (De Deyne et al., 2016) using a general principle that guides the organisation of the coarse-grained structure of semantic networks.

2. Experiment 1

In Experiment 1, adults sorted 140 familiar, nameable categories of objects. One group sorted pictures (e.g., a picture of a dog) and a second group sorted words (e.g., the word "dog"). In addition to investigating the nature of our stored semantic knowledge by examining whether, overall, taxonomic sorting dominated over thematic sorting, we also tested whether pictures were more likely than words to be sorted taxonomically. Greater taxonomic sorting of pictures could occur because visual similarity across items is more salient for pictures and because taxonomically related objects often look similar (Rosch et al., 1976). Supporting this hypothesis, Lin and Murphy (2001) found 17% less thematic matching when pictures were presented to adults in addition to words (see also Tare & Gelman, 2010). Alternatively, if people use a single, stable network of stored semantic knowledge then pictures and words should be sorted similarly.

In Experiment 1, unlike traditional small-scale sorting and matching tasks, the semantic relations that participants used to cluster items had to be inferred from the data rather than being specified a priori. This was done in two ways. First, we used mean subjective ratings of how strongly items in a cluster were related taxonomically and were related thematically, similar to measures that have been used in other tasks (Maguire, Brier, & Ferree, 2010; Mirman & Graziano, 2012). Second, to assess whether people sorted similarly to each other or idiosyncratically we developed two objective measures of categorization consistency based on Cramér's phi (Cramér, 1946). These novel measures assessed whether different people tested in the same condition showed

consistent sorting, and whether groups of people tested under different conditions differed in their sorting.

2.1. Method

2.1.1. Participants

Two groups of 15 adults (mean age 25, range 17–73) volunteered to take part in the study or participated for course credit.

2.1.2. Materials, design and procedure

Two sets of 140 cards were produced which comprised a physically and semantically diverse range of everyday objects taken from a wide range of superordinate categories. The picture set comprised 121 line drawings of familiar, nameable objects taken from Snodgrass and Vanderwart (1980) plus 19 objects drawn in a similar style. The word set comprised the names used by Snodgrass and Vanderwart (1980) for each of their objects (or the British English equivalent, e.g., "lorry" instead of truck) and names chosen by the first author for the remaining 19 items, see supplementary materials. The words were printed in a large font with the initial letter in uppercase and the remaining letters in lowercase. The stimuli were printed in black ink on white paper with a white border of up to 2 cm and then mounted onto cards. The picture stimuli ranged from 3 cm wide × 5 cm high to 12 × 9 cm and the word stimuli ranged from 4 × 3 cm to 9 × 5 cm.

Participants saw either the picture or the word cards, which were laid out upright and face-up on a table. A different card layout was used for each participant. Participants were told to sort the cards into clusters as they saw fit. There was no minimum size for a cluster but they had to sort all the items. No feedback was provided. Most participants finished sorting within 30 min but if they did not they were allowed to continue until they were finished. Sorting time was not recorded. Afterwards participants named each cluster which they had produced. Typical names provided were bathroom, nature, animals and tools. These names were used by the raters to help to generate S_{THEM} and S_{TAX} ratings, see Appendix A for details.

2.1.3. Dependent measures

The analyses for each of the three initial experiments in this paper first report objective measures of sorting consistency followed by subjective ratings of the structure of each cluster. The first set of analyses report the intra-group and inter-group consistency of sorting as assessed by the objective dependent measures of Φ_{INTRA} and Φ_{INTER} respectively. Φ_{INTRA} measures the consistency of clustering of items for participants in the same experimental group. This was indexed by means of a co-prediction metric (Cramér's phi; Cramér, 1946) that was computed for each pair of participants. Above-chance Φ_{INTRA} indicates that participants in a group agreed more than would be expected if they were using the same number of clusters as the observed participants but otherwise responded randomly. Φ_{INTER} provides an objective measure of agreement between experimental groups. Below-chance Φ_{INTER} indicates that participants agreed significantly less with each other than would be expected if the participants had been randomly allocated to the two experimental groups. This, in turn, means the participants in the two groups must have clustered the objects differently (see Appendix A for an extended explanation of this point). The question of whether Φ_{INTRA} and Φ_{INTER} were significantly different from chance was addressed by numerical methods, see Appendix A. See www.willslab.co.uk/phi/ for materials to support the use of these novel measures of sorting consistency in future research. The raw sorting data are also archived at this location with md5 checksum: 729f9d91bc0fa2166563ef3c67829845.[1]

The second set of analyses report analyses of variance (ANOVAs) using the subjective dependent measure of ratings of the structure of clusters. Five independent, blind raters used a scale ranging from 0 (no structure) to 9 (maximally coherent structure) to rate the organisation of each of the clusters produced by participants. See Appendix A for details. S_{TAX} is the mean across the five raters of their assessment of taxonomic structure and S_{THEM} is the corresponding measure for rated thematic structure. The subjective ratings allow us to identify whether people sorted using mainly taxonomic versus thematic relations.

2.2. Results

The mean number of clusters produced was 21. This was similar for the picture group (22; range 12–48 clusters) and the word group (21; range 14–28). There was a mean over participants of 7.1 items per cluster for both groups (range 1–49 items).

2.2.1. Objective measures of sorting consistency

First, both the picture group and the word group showed high intra-group agreement, as assessed using Φ_{INTRA}, see Table 1. Both groups were significantly above chance on this measure, so participants within each group agreed more with each other than would be expected if they had used the same number of clusters but otherwise had allocated items to clusters at random (see Appendix A). The absolute levels of agreement were comparable to previous large-set, free-sorting tasks (for example, the young adults in Haslam et al. (2007) were asked to free-sort a set of 60 Munsell colour chips and scored a mean Φ_{INTRA} of 0.78). Second, presentation format (pictures versus words) affected the clusters produced, as evidenced by a significantly below-chance value for Φ_{INTER} for the words vs. pictures comparison, see Table 2. This shows that pairs of picture and word participants agreed less with each other than would be expected for a random allocation of participants to the picture and word groups (see Appendix A). One possible reason for this difference relates to the small number of clusters produced which consisted of items with a single common perceptual feature. For example, these clusters might be labelled as long things, or as shiny things or small things. Such single property clusters may have been more common for the picture group (where perceptual features were more salient) and so this group may, therefore, have sorted more unidimensionally (Medin, Wattenmaker, & Hampson, 1987). This, in turn, could have influenced Φ_{INTER} without affecting Φ_{INTRA} (or S_{TAX} or S_{THEM}). Note that Φ_{INTRA} and Φ_{INTER} are different dependent variables measuring different things using different methods of calculation, and thus have different chance levels (see Appendix A for details).

2.2.2. Subjective ratings of taxonomic versus thematic structure

2.2.2.1. By participants. An ANOVA was conducted with one within-subjects factor of ratings of cluster structure (S_{TAX} vs. S_{THEM}) and one between-subjects factor of stimulus type (pictures vs. words). People's clusters were rated as having significantly greater thematic structure (mean S_{THEM} = 4.97; range 3.9 to 5.8) than taxonomic structure (mean S_{TAX} = 3.56; range 2.8 to 4.4), $F(1,28) = 117.29$, $p < 0.001$, partial

Table 1
Objective measures of intra-group consistency as indexed by Cramér's phi (Φ_{INTRA}) for each group in the four free-sorting experiments reported here. For all groups, Φ_{INTRA} was significantly higher than would be expected if participants were producing clusters randomly (see Appendix A for further details).

Experiment + Group	Φ_{INTRA}	p
E1 Pictures	0.79	<0.01
E1 Words	0.77	<0.01
E2 Thematic-primed	0.74	<0.01
E2 Control-unprimed	0.72	<0.01
E2 Taxonomic-primed	0.72	<0.01
E3 Thematic Instructions	0.64	<0.01
E3 Taxonomic Instructions	0.76	<0.01
E4 Thematic Instructions	0.72	<0.01
E4 Taxonomic Instructions	0.82	<0.01

[1] Publication of an MD5 checksum allows the reader to independently confirm that the raw data in the archive is unchanged.

Table 2
Objective measures of inter-group consistency as indexed by Cramér's phi (Φ_{INTER}) for all pairs of groups of interest. For three of the four comparisons, consistency was significantly lower than that produced by a random allocation of participants to groups (see Appendix A for further details), indicating that the two groups clustered stimuli differently to each other.

Pairs of (Experiment + Group)	Φ_{INTER}	p
E1 Pictures, E1 Words	0.77	<0.01
E2 Thematic-primed, E2 Taxonomic-primed	0.73	n.s.
E3 Thematic Instructions, E3 Taxonomic Instructions	0.67	<0.01
E4 Thematic Instructions, E4 Taxonomic Instructions	0.72	<0.05

$\eta^2 = 0.81$. People did not, though, use one type of semantic relation exclusively: no participant produced clusters with both very high S_{THEM} ratings and very low S_{TAX} ratings or vice versa: the range of [S_{THEM}–S_{TAX}] for participants was +2.4 down to −0.02.

Stimulus type (pictures vs. words) did not significantly affect overall ratings, $F(1,28) = 0.00$, $p = 0.9$, partial $\eta^2 = 0.00$, and the interaction between stimulus type and structure ratings (thematic vs. taxonomic) was also not significant, $F(1,28) = 0.36$, $p = 0.5$, partial $\eta^2 = 0.01$, see Fig. 1. Thus, there was no evidence that the difference between the picture and word groups revealed by the below-chance Φ_{INTER} was due to the word group making more use of thematic clustering.

We checked whether the results of this initial ANOVA changed when each cluster was given equal weight (regardless of the number of items in it) when calculating S_{TAX} and S_{THEM} (see Appendix A for further details). This second ANOVA replicated the pattern of results of the first so our findings were not sensitive to the method of calculation used.

2.2.2.2. By items. We examined whether the results found across participants generalised across the 140 items. For each item we took the mean of the S_{THEM} ratings for the 30 clusters which contained that item (each participant contributed one cluster; we ignored whether a given participant sorted pictures or words). Mean S_{TAX} ratings were calculated in the same way. For around one third of the items (45 of 140), that item was placed in clusters rated as having more taxonomic than thematic structure, so [$S_{TAX} > S_{THEM}$]. In comparison, in the by participants analysis reported above, only one of the 30 participants made clusters which, on average, were rated as having more taxonomic than thematic structure. There was also a greater range for S_{THEM} (2.9 to 6.7) and, especially, S_{TAX} (0.6 to 7.7) for items than for participants, with [S_{THEM} - S_{TAX}] fourfold greater for items (+5.6 to −4.3) than for participants (+2.4 to −0.02). Thus, compared to individual participants, individual items were more likely to be consistently sorted into either mainly taxonomically organised clusters or mainly thematically organised clusters. Sorting by participants showed less variability across individuals and nobody sorted overwhelmingly taxonomically.

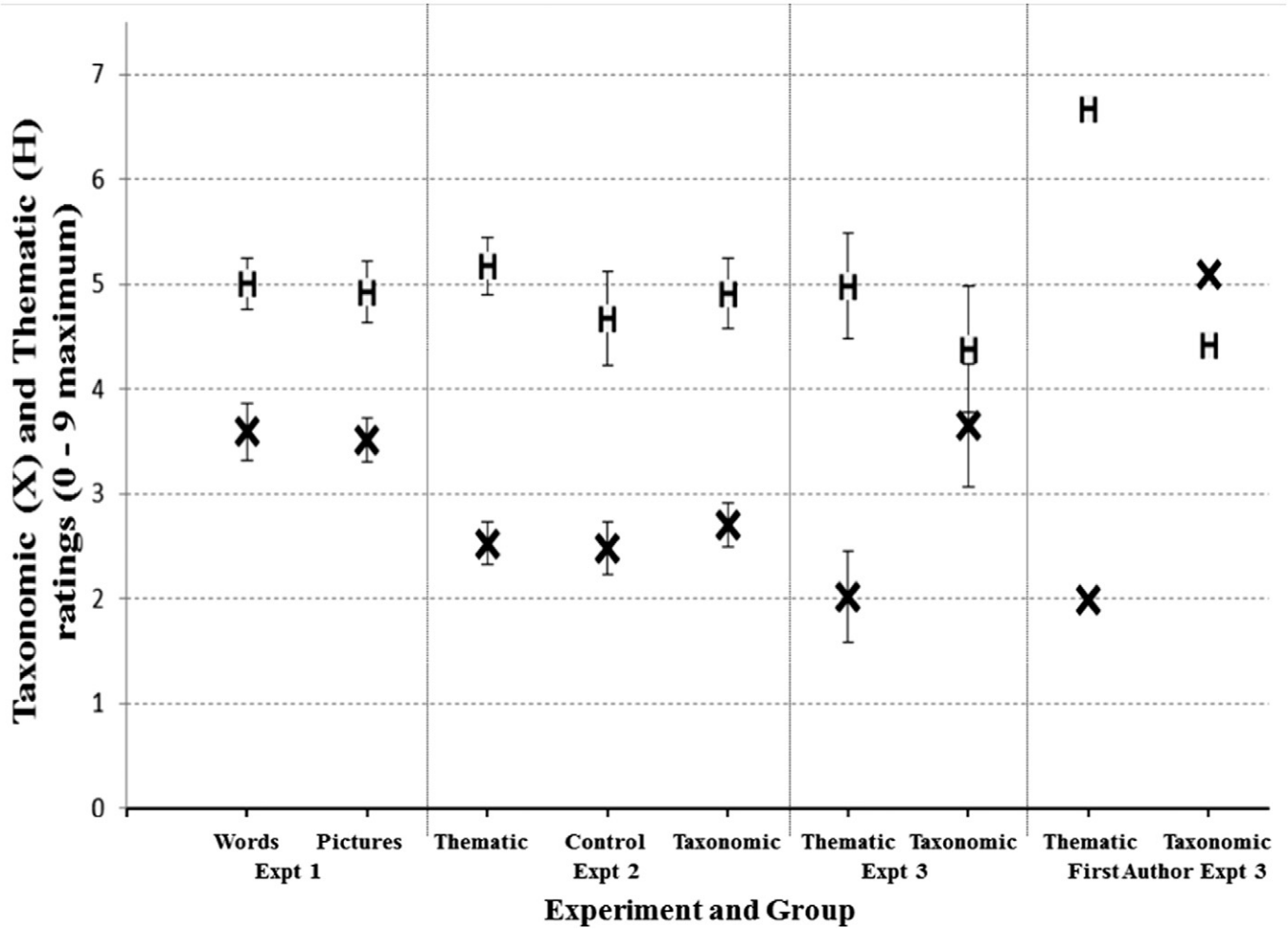

Fig. 1. Mean subjective ratings of the sorted clusters produced by participants based on thematic structure ($S_{THEM;}$ H markers) and taxonomic structure ($S_{TAX;}$ X markers) on a scale from 0 (no structure to the cluster, i.e. a random collection of items) to 9 (maximally coherent structure). These results show ratings when the sorting of each item was weighted equally. The word and picture group in Experiment 1 were shown different types of stimuli. The thematic-primed, control-unprimed and taxonomic-primed groups in Experiment 2 were primed to sort differently prior to doing the main sorting task. The thematic-instructed and taxonomic-instructed groups in Experiment 3 were explicitly told to sort thematically and taxonomically respectively. Error bars show 95% confidence intervals for between-subjects variation across groups.

This items analysis was repeated but taking stimulus type into account. Here, the S_{THEM} and S_{TAX} ratings for a given item were averaged over the 15 clusters that each word group participant sorted that item into and, separately, for the 15 clusters that each picture group participant sorted that item into. There were high correlations between these word and picture ratings for both S_{TAX}, $r(140) = +0.93$, $p < 0.001$, and for S_{THEM}, $r(140) = +0.83$, $p < 0.001$. This supports the by-participants analysis in indicating that the word and picture groups used similar semantic relations to sort a given item.

2.3. Discussion

Our first finding from Experiment 1 was that when people freely sorted 140 familiar objects they did not produce idiosyncratic solutions. Instead, our objective measure, Φ_{INTRA}, indicated that they agreed with each other quite well about how to cluster items - about as much as a similar population of adults agreed about how to sort colours (Haslam et al., 2007), see Table 1.

Second, people made clusters that, overall, were subjectively rated as having more thematic than taxonomic structure, see Fig. 1. Thus people mainly clustered items because they co-occurred in common situations, locations or times, rather than clustering them taxonomically due to their perceptual or functional similarities. This suggests that there were more thematic than taxonomic relations between these objects available in stored semantic knowledge. Nevertheless, despite the overall dominance of thematically organised clusters, every participant produced clusters of objects that were rated as having at least moderate taxonomic structure. Thus people did not exclusively use stored, episodic and associative information to structure their clusters; everyone also used taxonomic relations. The consistency of people's sorting was investigated further in Experiment 2.

Third, presentation format (words versus pictures) had no detectable effect on subjective ratings of the relative prevalence of taxonomic and thematic sorting, see Fig. 1. In particular, there was no evidence that people's sorting relied more on visual information for the pictures, leading to more taxonomically organised clusters in the picture group compared to the word group. In the by-participants analyses, the picture group clusters were not rated as more taxonomically organised than those of the word group. Furthermore, in the by-items analysis there was a high correlation (+0.93) between the taxonomic ratings of the clusters made by the word group and the picture group for a given item, and also a high correlation (+0.83) for thematic ratings. Thus participants usually used the same type of semantic relation to sort a given item irrespective of whether it was referred to using a picture or its name. This suggests that people consistently accessed the same type of information from their stored semantic networks irrespective of the presentation format used to access those networks.

3. Experiment 2

All but one participant in Experiment 1 sorted the 140 items into clusters rated as having more thematic than taxonomic organisation. Nevertheless, despite the dominant use of thematic relations, everyone produced clusters with some taxonomic organisation. Experiment 2 probed the generality of these two findings by, first, using a different stimulus set, second, by examining the effect of sorting order, and third, by manipulating priming. Regarding the first issue, the main sorting task in Experiment 2 used a new set of 150 everyday objects. The objects were selected to be diverse, yet familiar, and were represented by coloured pictures.

Second, people may have used a mix of thematic and taxonomic relations in Experiment 1 because they could not sort all of the items using a single type of semantic knowledge. In particular, participants may have wanted to sort purely taxonomically but have been unable to sort all of the items in this way. In order to investigate this possibility, in Experiment 2 we recorded the order in which clusters were produced. We reasoned that if participants wanted to sort taxonomically their initial clusters should be organised taxonomically. They should only switch to using other types of semantic relations to organise their clusters when they were unable to easily create more taxonomic clusters. In contrast, if maintaining sorting consistency was not a priority then people should produce a mix of taxonomic and thematic clusters right from the start of sorting.

Third, a different aspect of sorting persistence was tested in Experiment 2. We investigated whether sorting was determined solely by the objects presented or whether people could be primed to use a specific type of semantic relation (taxonomic versus thematic). Milton and Wills (2009) reported that priming could influence a person's subsequent sorting. They showed stimuli that encouraged unidimensional sorting to one group and they showed stimuli that encouraged family resemblance sorting based on overall similarity to another group. On subsequent sorts, even a week later, people tended to sort as they had done during priming (whether unidimensional or family resemblance) even when different stimuli were shown and no feedback was provided. This sorting persistence is consistent with evidence from small-scale, free-sorting and matching tasks which also indicate that people may prefer to sort consistently (Koriat & Melkman, 1981; Lin & Murphy, 2001; Murphy, 2001; Simmons & Estes, 2008; Wattenmaker, 1995).

A similar priming method to that used by Milton and Wills (2009) was employed in Experiment 2 to encourage the use of different types of semantic relations to organise sorting across three groups. The thematic-primed and taxonomic-primed groups each did different priming tasks. These tasks were followed by the main task in which all three groups sorted the same set of 150 familiar, nameable objects (this was the only task for the third, unprimed-control group). The priming sort used stimuli which clustered naturally into three sets of three objects. The thematic-primed group sorted items which readily clustered thematically but which could not easily be sorted taxonomically. The taxonomic-primed group sorted items which could easily be clustered taxonomically but not thematically. The thematic-primed and taxonomic-primed groups were therefore expected to do the priming sort thematically and taxonomically respectively. If, after priming, people persist in using the same type of semantic relations to cluster items in the subsequent main sorting task then the thematic-primed group should sort more thematically and the taxonomic-primed group should sort more taxonomically compared to the unprimed-control group. In contrast, if sorting mainly reflects the identity of the objects used in the main sorting task then sorting should be similar across all three groups.

3.1. Method

3.1.1. Participants

Three groups of 17 volunteers (mean age 21, range 18–63) took part in the study for course credit.

3.1.2. Materials and apparatus

The sorting cards were all coloured pictures of familiar, nameable objects which were chosen to come from a physically and semantically diverse set of basic level categories. The images on each card were a mix of photographs of real objects and more schematic pictures. These images were scaled to approximately 4 × 4 cm and were then printed in colour on white paper. The pictures were then glued and laminated onto 7 × 7 cm white cards. Most cards showed a single object in isolation but some showed multiple examples of the object (e.g., a bunch of bananas, several coins). Three people named all of the objects and any objects that were not consistently and correctly named were replaced.

Three sets of cards were used: taxonomic prime cards, thematic prime cards and the main sort cards. The nine taxonomic prime cards comprised trios of items from different superordinate categories (animals - cat, dog, rabbit; fruit - grapes, pineapple, strawberry; and vehicles

- aeroplane, car, train). Similarly, the nine thematic prime cards comprised trios of items from different themes (office - notepad, pen, stapler; dog - metal water bowl, dog, kennel; and beach - bucket with spade, deckchair, ice-cream cone). Finally, there were 150 main sort cards of which 53 had the same labels as the stimuli used in Experiment 1, see supplementary materials.

3.1.3. Design and procedure

The taxonomic-primed and thematic-primed groups started by doing a quick priming sort. As a cover story they were told that this was to allow them to practice sorting before they did the main sorting task. The taxonomic-primed group were shown the nine taxonomic prime cards and the thematic-primed group were shown the nine thematic prime cards. The cards were arranged in a 3 × 3 matrix with no row or column including all three taxonomically or thematically or related items. Participants were given written instructions adapted from Murphy (2001) which told them to move the cards into the clusters that seemed most natural to them. No feedback was provided but all participants produced the expected taxonomic or thematic clusters. This task usually took 1–2 min to complete.

All three groups then did the main sorting task which was the only task for the control-unprimed group. Participants were shown the 150 cards arranged upright on a table in a 10 × 15 matrix. They were given the same written instructions as in the priming sort task except that they were also told that they could move an item from one cluster to another at any time if they changed their mind. They were allowed to make as many clusters as they wished provided that they had at least two items in every cluster and that every item was placed in a cluster. Participants were asked to make their first cluster on their far left and to make each successive cluster to the right of the previous cluster. This meant that the order in which clusters were started could be recorded. No feedback was provided. After the main sort participants provided a name for each of the clusters that they had created. As in Experiment 1 these names were used by the raters to help to generate S_{THEM} and S_{TAX} ratings, see Appendix A for details.

3.2. Results

The mean time to complete the main sorting task was 20 min (range 12–41 min; means of 19, 19 and 22 min for the control-unprimed, thematic-primed and taxonomic-primed groups). The mean number of clusters produced was 21 with similar numbers for the three groups (range 10–34; means of 19, 21 and 22 for the control-unprimed, thematic-primed and taxonomic-primed groups). There was a mean over participants of 8.0 items per cluster (range 2–50; means of 9.2, 7.3 and 7.7 for the control-unprimed, thematic-primed and taxonomic-primed groups). Overall sorting performance for these 150 items was thus similar across the three groups and it was also similar to Experiment 1, where there was a mean of 21 clusters and 6.6 items per cluster for a 140 item sort.

3.2.1. Objective measures of sorting consistency

First, all three groups showed high intra-group agreement, as assessed using Φ_{INTRA}, see Table 1. All groups were significantly above chance on Φ_{INTRA}. Thus people again generally agreed with each other about how to cluster this large set of items, suggesting that they were using similar information to sort so they were not doing the task idiosyncratically. The absolute levels of agreement on Φ_{INTRA} were comparable to those of Experiment 1. Second, people sorted similarly regardless of the nature of the priming task. Priming (taxonomic versus thematic) did not significantly change the consistency of the clusters produced, with Φ_{INTER} not significantly different from chance, see Table 2.

3.2.2. Subjective ratings of taxonomic versus thematic structure

3.2.2.1. By participants. An ANOVA revealed that, as in Experiment 1, people's clusters were rated as having more thematic structure (mean $S_{THEM} = 4.92$, range 3.1 to 6.0) than taxonomic structure (mean $S_{TAX} = 2.57$, range 1.4 to 3.3), $F(1,48) = 731.80$, $p < 0.001$, partial $\eta^2 = 0.94$. Priming (thematic, control or taxonomic) did not significantly affect overall ratings, $F(2,48) = 1.52$, $p = 0.2$, partial $\eta^2 = 0.06$. Finally, the interaction between priming and structure ratings (thematic vs. taxonomic) was marginally significant, $F(2,48) = 2.95$, $p = 0.06$, partial $\eta^2 = 0.11$, see Fig. 1. There was a trend in the predicted direction with somewhat greater ratings of thematic structure for the thematic-primed group ($S_{THEM} = 5.2$, $S_{TAX} = 2.5$, a difference of +2.6) than the control-unprimed group (4.7 and 2.5; a difference of +2.2) and the taxonomic-primed group (4.9 and 2.7; a difference of +2.2). This effect was, though, weak. Even in the taxonomic-primed group [$S_{THEM} > S_{TAX}$] for every participant. Thus nobody in this group made predominantly taxonomic clusters in the main sorting task even though they had sorted the nine prime items taxonomically immediately before starting the main sort.

We checked whether the results of this initial ANOVA changed when each cluster was given equal weight (regardless of the number of items in it) when calculating S_{TAX} and S_{THEM} (see Appendix A for further details). This second ANOVA replicated the pattern of results of the first so our findings were not sensitive to the method of calculation used.

3.2.2.2. By items. We examined whether the results found across participants generalised across the 150 items by calculating the mean S_{THEM} and S_{TAX} ratings for a given item over the 51 clusters (one per participant) which contained that item. In the by participants analysis reported above, none of the 51 participants produced clusters rated as having more taxonomic than thematic structure (so with [$S_{TAX} > S_{THEM}$]), whereas 35 of the 150 items had [$S_{TAX} > S_{THEM}$]. Thus, as in Experiment 1, a substantial minority of individual items were sorted mainly taxonomically whereas all participants used mainly thematic sorting.

3.2.2.3. Cluster order analysis. This analysis examined whether people changed how they clustered over the course of their sort. If people had wanted to sort exclusively using taxonomic relations then they would be expected to begin by sorting their first few clusters taxonomically even if they were later forced to switch to making thematically organised clusters when they could no longer easily create taxonomic clusters. There was no evidence to support the strong version of this hypothesis. Even for the first cluster produced, where S_{TAX} was at its greatest, mean S_{THEM} (4.5) ratings were similar to S_{TAX} (4.6) ratings. Taxonomic relations thus never dominated clustering.

There was, though, robust evidence for a weaker version of this hypothesis because people's clusters became more thematically organised and less taxonomically organised over the course of their sort. For the first 24 clusters produced (where at least 20 participants provided data for each cluster position), there was a strong, positive correlation between S_{THEM} and the position in the sort that the cluster was produced, $r(24) = +0.60$, $p = 0.003$, and a strong negative correlation between S_{TAX} and cluster position, $r(24) = -0.67$, $p < 0.001$.

3.2.2.4. The distribution of cluster ratings. This analysis was based on the subjective S_{THEM} and S_{TAX} ratings from the 1695 clusters produced by the 81 participants tested in Experiments 1 and 2. These ratings, which were on a scale from 0 (no structure) to 9 (maximally coherent structure), were divided into three bands. Mean S_{THEM} and S_{TAX} ratings of 6 and above were classed as having strong structure, ratings < 4 were classed as having weak structure and the remaining ratings were classed as having medium structure, see Fig. 2. Overall, 36% of clusters had strong thematic structure whereas only half as many, 16%, had strong taxonomic structure. Conversely, only 26% of clusters had weak thematic structure whereas twice as many, 60%, had weak taxonomic

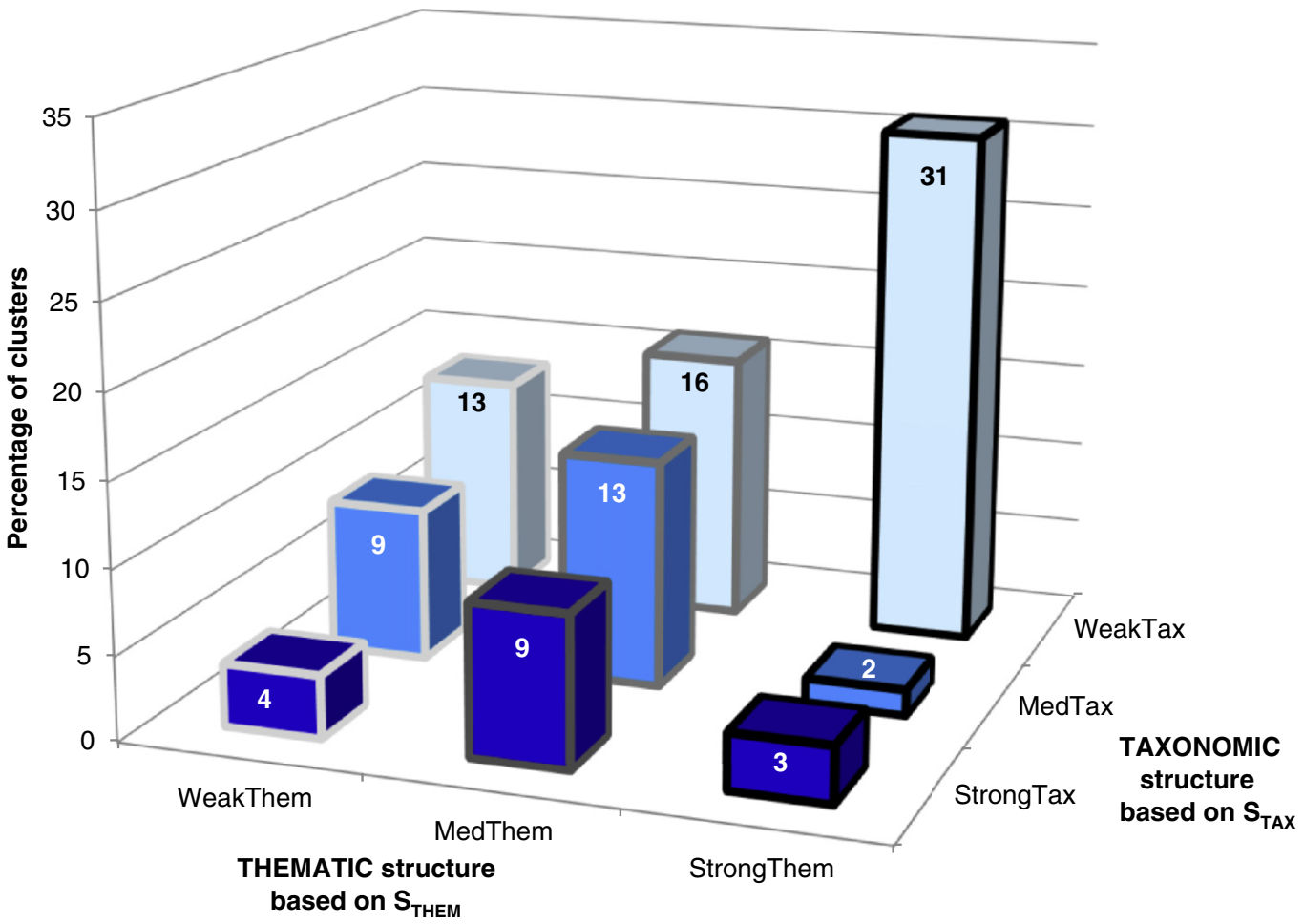

Fig. 2. Percentages of the 1695 clusters produced by the 81 participants in Experiments 1 and 2 which were rated as having strong, medium or weak thematic structure (based on S_{THEM}) and strong, medium or weak taxonomic structure (based on S_{TAX}). There were fewer items per cluster for more strongly structured clusters for both thematic clusters (with means of 4.3, 7.0 and 7.8 items for strong, medium and weak clusters) and taxonomic clusters (means of 4.3, 6.9 and 7.9 items respectively).

structure. Many clusters (31%) were only organised thematically: they were rated as having strong thematic and weak taxonomic structure. Examples of such clusters with just two members included toothbrush + toothpaste; purse + coins; and post box + envelope. In contrast, eight times fewer clusters (just 4%) were organised primarily taxonomically, with weak thematic and strong taxonomic ratings. Two member examples included pool-table + table; parrot + duck; and broom + paintbrush.

3.3. Discussion

The results from Experiment 2 replicated the main findings from Experiment 1 with a different set of objects. This argues against the suggestion that the results of Experiment 1 resulted from an unrepresentative choice of items. As in Experiment 1, the objective measure of sorting consistency, Φ_{INTRA}, showed that participants generally agreed with each other about how these items should be clustered, see Table 1, so there was a systematic basis to their sorting behaviour. Again replicating Experiment 1, thematic relations dominated performance: every participant produced clusters rated as having more thematic than taxonomic structure. This suggests that for the 150 everyday objects tested in Experiment 2, as for the 140 objects tested in Experiment 1, there were more thematic relations (such as common situations, locations or times) available as a basis for sorting, relative to taxonomic relations involving perceptual or functional similarity. It is, though, also important to emphasise that the results of Experiment 2, like those of Experiment 1, clearly show that taxonomic information is an important component of our semantic knowledge. Both types of semantic relations were used to sort, see Figs. 1 and 2, and nearly a quarter of the items were mainly sorted taxonomically.

The cluster order analysis revealed that the semantic relations used to sort altered as the task progressed. The first cluster of items produced had similar ratings of S_{THEM} and S_{TAX} but subsequent clusters were rated as being increasingly thematically structured and decreasingly taxonomically structured. Koriat and Melkman (1981) found similar results using clustering in a word recall task. These changes to the nature of clustering as the sort progressed are consistent with taxonomic relations being as salient, but less common, than thematic relations such that a few taxonomic clusters could easily be created at the start of a sort but that producing taxonomic clusters would become harder as the sort progressed. Importantly, though, even at the start of sorting taxonomic clusters did not dominate. Our results thus provide no support for the claim that participants tried to sort consistently using only taxonomic relations. It could be argued that, after their initial inspection of the items, participants realised that they could not sort all of the items taxonomically and so they did not attempt to do so. We think that this explanation is unlikely because there were so many items that people could not retain them all simultaneously in working memory. Nevertheless, it would be useful to confirm this in future research by monitoring people's eye movements during sorting in order to try to infer the information that they used to plan their sort. The influence of pre-planning on clustering could also be investigated directly, by repeating this

study but giving participants only one item at a time to add to clusters and not informing them about the total number of items.

Finally, the priming manipulation had little influence on sorting. The thematic-primed group did not produce significantly more thematically structured clusters than the control-unprimed group and the taxonomic-primed group. For the taxonomic-primed group, during priming participants produced taxonomically organised clusters. However, this failed to then overturn the dominance of thematic sorting: in the subsequent main sort every participant still produced clusters that were rated as more thematically than taxonomically organised. Finally, the objective measure of inter-group agreement (Φ_{INTER}) did not reveal a difference in the sorting consistency of the thematic-primed versus taxonomic-primed groups, see Table 2.

One reason for the lack of an effect of priming in Experiment 2 could have been that it was difficult to produce many clusters with a pure taxonomic (or, indeed, a pure thematic) organisation with this stimulus set. This observation led us to conduct Experiment 3. In this experiment people were explicitly directed to sort either taxonomically or thematically in order to assess the extent to which they could exclusively use one type of semantic relation to sort these stimuli.

4. Experiment 3

Our results so far show that, in large-scale sorting tasks presenting familiar, everyday objects, 80 of the 81 people tested produced clusters rated as having more thematic than taxonomic structure. Very few of these clusters (4%) were rated as having weak thematic but strong taxonomic structure, whilst many more (31%) were rated as having strong thematic but weak taxonomic structure (see Fig. 2). This suggests that we store more thematic than taxonomic relations in our long-term semantic networks. However, an alternative possibility is that sorting in Experiments 1 and 2 was not driven by the number of stored semantic relations *available* between the objects being sorted but was, instead, determined by people *choosing* to sort thematically rather than taxonomically. This latter hypothesis predicts that people could have sorted the objects in Experiments 1 and 2 more purely taxonomically but they chose not to.

To test this possibility, participants in Experiment 3 sorted the same set of 150 objects as in Experiment 2 but they were explicitly instructed to sort using either taxonomic or thematic relations. If the taxonomic-instructed group in Experiment 3 successfully sorted all of the items into well-structured, taxonomic clusters this would show that in Experiment 2 participants could have sorted exclusively taxonomically if they had wished to do so. This, in turn, would indicate that the dominance of thematic sorting in Experiments 1 and 2 resulted from people *choosing* to organise their clusters thematically rather than taxonomically. If, though, the taxonomic-instructed group in Experiment 3 could not create many taxonomically organised clusters and, instead, sorted like the thematic-instructed group (and like the free-sorting participants in Experiments 1 and 2) this would show that there were relatively few taxonomic relations *available*. Thus Experiment 3 tested whether the prevalence of thematic sorting found so far primarily reflected people *preferring* to use thematic relations or whether it reflected the greater number of thematic compared to taxonomic relations *available*. Only in the former case should sorting be sensitive to instructions.

4.1. Method

4.1.1. Participants

Twenty-five volunteers (13 instructed taxonomically and 12 instructed thematically; mean age 19; range 18–24) took part in the study for course credit.

4.1.2. Materials and apparatus

These were identical to Experiment 2.

4.1.3. Design and procedure

The main sorting task was similar to Experiment 2 except that, before it began, the difference between taxonomic and thematic categories was explained to participants and the thematic-instructed group was told to sort thematically whilst the taxonomic-instructed group was told to sort taxonomically. Both groups were given definitions based on those used by Murphy (2001). Taxonomic clusters were said to comprise "items that are in the same category, such as furniture or mammals, or at least that are similar in some respect" whilst thematic clusters included "things that are grouped together because they occur in the same setting or event or because one of them fulfils a function of the other one". All participants also did both the taxonomic and the thematic nine card priming sort task used in Experiment 2 to concretely illustrate to them the difference between taxonomic and thematic sorting. The taxonomic-instructed group was then told to do the main sort producing only taxonomic clusters whilst the thematic-instructed group was told to produce only thematic clusters. They were told not to use both types of semantic relation to cluster. To emphasise this instruction the cards from the priming sort task were left visible with notices saying "sort like this" and "not like this" placed above them.

4.2. Results

The mean time to complete the main sorting task was 19 min (range 15–23 min). The mean number of clusters produced (21 and 28 for the thematic and taxonomic groups respectively, with a range of 11–50 clusters) was similar to Experiment 2, as was the mean over participants of items per cluster (7.6 and 6.3 for the thematic and taxonomic groups; range 2–34 items).

4.2.1. Objective measures of sorting consistency

Both groups showed high intra-group agreement, as assessed using Φ_{INTRA}, see Table 1, with sorting in both groups being significantly above chance, indicating that participants did not sort idiosyncratically in either group. Sorting instructions (taxonomic versus thematic) affected the level of agreement in how participants sorted, with Φ_{INTER} significantly less than chance, see Table 2. Thus sorting instructions influenced this measure of inter-group consistency.

4.2.2. Subjective ratings of taxonomic versus thematic structure

4.2.2.1. By participants. An ANOVA revealed a significant difference between thematic and taxonomic ratings of cluster structure (S_{THEM} vs. S_{TAX}; $F(1,23) = 49.43$, $p < 0.001$, partial $\eta^2 = 0.68$), and also of thematic versus taxonomic sorting instructions ($F(1,23) = 5.00$, $p = 0.03$, partial $\eta^2 = 0.18$). Finally, the interaction between structure ratings and sorting instructions was significant ($F(1,23) = 18.12$, $p < 0.001$, partial $\eta^2 = 0.44$, see Fig. 1). Bonferroni adjusted pairwise comparisons for the *thematic-instructed group* showed that their clusters had more thematic structure (mean $S_{THEM} = 4.98$; range 3.1 to 5.9) than taxonomic structure (mean $S_{TAX} = 2.02$; range 1.3 to 3.3), $p < 0.001$. In contrast, the *taxonomic-instructed group* had no significant difference between ratings of S_{THEM} (mean 4.38; range 1.6–5.3) and S_{TAX} (mean 3.65; range 1.9–4.8), $p = 0.06$. Furthermore, note that the trend for this group was to sort more thematically than taxonomically, contrary to their instructions.

We checked whether the results of this initial ANOVA changed when each cluster was given equal weight (regardless of the number of items in it) when calculating S_{TAX} and S_{THEM} (see Appendix A for further details). This second ANOVA replicated the pattern of results of the first so our findings were not sensitive to the method of calculation used.

4.2.2.2. By items. We examined whether the results found across participants generalised across items by calculating the mean S_{THEM} and S_{TAX} ratings for each item for the 13 clusters (one for each of the

taxonomically-instructed participants) which contained that item and, separately, for the 12 clusters (one for each of the thematically-instructed participants) which included that item. In Experiments 1 and 2 we found mainly taxonomic sorting for a substantial minority of items ([$S_{THEM} < S_{TAX}$] for 45/140 and 35/150 items respectively). In Experiment 3, for the *taxonomic-instructed group*, [$S_{THEM} < S_{TAX}$] for 61/150 items (mean +0.7, range +4.1 to −3.0). In contrast, none of the 150 items sorted by the *thematic-instructed group* in Experiment 3 had [$S_{THEM} < S_{TAX}$] (mean +3.0, range +6.0 to +0.2). Thus every item could be placed into clusters rated as mainly thematic (by the *thematic-instructed group*) whereas less than half of the items could be sorted into clusters rated as mainly taxonomic (by the *taxonomic-instructed group*).

4.2.2.3. First author sorting results. The naive participants tested in Experiment 3 may not have fully understood the difference between thematic and taxonomic sorting or they may have ignored their instructions. If so then the results reported above may have underestimated the extent to which these stimuli could be sorted exclusively thematically or taxonomically. To address this possibility the first author sorted the 150 stimuli twice, once thematically and once taxonomically. For the thematic sort, S_{THEM} (7.0) was higher than for any participant in Experiment 3 and S_{TAX} was just 1.7, see Fig. 1.[2] For the taxonomic sort, S_{THEM} was 4.4 whilst S_{TAX} (5.4) was higher than for any participant in Experiment 3, see Fig. 1. Removing the first author's ratings to calculate these ratings produced similar results (thematic sort: 6.9 and 1.5, taxonomic sort: 4.4 and 5.1, for S_{THEM} and S_{TAX} respectively). The first author thus sorted both more thematically and more taxonomically than the naive sorters tested in Experiment 3. Nevertheless, just like the naive sorters, she was more effective at using purely thematic rather than purely taxonomic relations to organise her clusters.

4.3. Discussion

The results of Experiment 3 fell between the two predictions that we made about whether the thematic dominance found in Experiments 1 and 2 was due to participants *choosing* to sort thematically or was due to the greater number of thematic relations *available*. Explicit instructions did significantly alter sorting in Experiment 3 as measured both objectively, by Φ_{INTER} (see Table 2), and subjectively, by S_{THEM} and S_{TAX} (see Fig. 1). Thus people could, to some extent, *choose* which semantic relations they used to sort. However, this effect was modest. The thematic-instructed group did produce clusters rated as more thematically than taxonomically organised but their taxonomic ratings were not at floor (see Fig. 1) whilst the clusters produced by the taxonomic-instructed participants were rated as having similar levels of taxonomic and thematic structure. These results are consistent with the conclusion from Experiments 1 and 2 that there are more thematic than taxonomic relations *available* in our stored, long-term semantic knowledge to organise clusters in a large-scale sorting task.

It might be proposed that, in Experiment 3, the taxonomic-instructed group decreased their criterion level for producing a cluster relative to the free-sorting participants in Experiments 1 and 2. If so, then the results of these different groups could not be meaningfully compared. However, if the taxonomic-instructed group had weakened their criteria then this should have reduced S_{TAX}, S_{THEM} and Φ_{INTRA}. Contrary to this prediction, this group produced clusters with higher ratings of S_{TAX} than any other group who sorted these stimuli, see Fig. 1. Note that S_{TAX} and S_{THEM} were taken from independent, subjective ratings of the coherence of a cluster, with raters blind to which clusters came from which condition or which experiment. The objective measure of sorting, Φ_{INTRA}, also demonstrated that the taxonomic-instructed group showed high intra-group consistency (see Table 1). There was therefore no evidence from either independent, subjective or objective measures that suggested that the taxonomic-instructed group in Experiment 3 used weaker criteria for creating clusters.

5. Experiment 4

In Experiment 3, it could be argued that the clusters made by the taxonomic-instructed group did not produce high S_{TAX} ratings because many items could not be sorted taxonomically and yet participants were forced to place every item into a cluster. This concern was addressed in Experiment 4, which largely replicated Experiment 3 but participants were allowed to discard items which they found difficult to sort into a random pile. This enabled us to estimate what proportion of the items participants believed could not be sorted as instructed. If the taxonomic-instructed group in Experiment 3 had struggled to sort more of the items than the thematic-instructed group then, in Experiment 4, the taxonomic-instructed group should discard more items into a random pile than the thematic-instructed group. This account also predicts that any increase in ratings in Experiment 4, compared to Experiment 3, should be greater for S_{TAX} for the taxonomic-instructed group than for S_{THEM} for the thematic-instructed group, because the taxonomic-instructed group should benefit more from being able to discard items into a random pile.

5.1. Method

5.1.1. Participants

Thirty-four volunteers (17 instructed taxonomically and 17 instructed thematically; mean age 20; range 18–26) took part in the study for course credit.

5.1.2. Materials and apparatus

These were identical to Experiment 3.

5.1.3. Design and procedure

The main sorting task was similar to Experiment 3[3] except that in the main sort the participants were told that they did not need to sort every stimulus into a cluster, though they were encouraged to try to sort as many as possible. They were told to leave any unsorted stimuli in a random pile at the end.

5.2. Results

The mean time to complete the main sorting task was 20 min (range 18–26 min). The mean number of clusters produced (20 and 23 for the thematic-instructed and taxonomic-instructed groups respectively, with a range of 7–39 clusters) was similar to Experiment 3, as was the mean over participants of items per cluster (9.2 and 7.1 for the thematic-instructed and taxonomic-instructed groups respectively; range 2–72 items).

5.2.1. Objective measures of sorting consistency

Items placed in the random piles were treated as not being clustered together in these analyses. Both groups showed high intra-group agreement, as assessed using Φ_{INTRA}, see Table 1, with sorting in both groups being significantly above chance, indicating that participants did not sort idiosyncratically in either group. Sorting instructions (taxonomic

[2] These two sorts were conducted after ratings of the taxonomic and thematic organisation of clusters had been collected. However, many of the first author's clusters were identical to clusters that had been rated (24/42 of the thematic clusters and 14/39 of the taxonomic clusters). Here, the mean rating provided for these identical clusters was used. For the remaining clusters, the rating for the most similar cluster was used; in most cases these shared all but one or two items with the first author's clusters.

[3] A further difference in Experiment 4 was that the main sorting task was preceded by participants being given two minutes to memorise the 150 stimuli. They then had to freely recall the names of as many of the stimuli as possible. This free recall task was repeated after completion of the main sort.

versus thematic) affected the level of agreement in how participants sorted, with Φ_{INTER} significantly less than chance, see Table 2. Thus sorting instructions influenced this measure of inter-group consistency.

5.2.2. Random piles

Most participants produced a random pile (16/17 of the taxonomic-instructed group and 14/17 of the thematic-instructed group). There was considerable variability in the size of the random piles produced (range 5–72 and 4–66 for the taxonomic-instructed and thematic-instructed groups respectively). Including those participants who did not make a random pile, the mean number of items in the random pile was 28 and 34 for the taxonomic-instructed and thematic-instructed groups respectively. An independent samples *t*-test found no evidence that the two groups differed in the size of their random piles, $t(32) = 0.790$, $p = 0.4$. In summary, the proportion of items discarded varied considerably, between none and nearly half of the items in both groups, and there was no evidence that the taxonomic-instructed group chose to discard more items than the thematic-instructed group.

5.2.3. Subjective ratings of taxonomic versus thematic structure

5.2.3.1. By participants. An ANOVA revealed a significant difference between thematic and taxonomic ratings of cluster structure (S_{THEM} vs. S_{TAX}; $F(1,32) = 96.01$, $p < 0.001$, partial $\eta^2 = 0.75$), and also of thematic versus taxonomic sorting instructions ($F(1,32) = 23.91$, $p < 0.001$, partial $\eta^2 = 0.43$). Finally, the interaction between structure ratings and sorting instructions was significant ($F(1,32) = 112.39$, $p < 0.001$, partial $\eta^2 = 0.78$). Bonferroni adjusted pairwise comparisons for the *thematic-instructed group* showed that their clusters had more thematic structure (mean $S_{THEM} = 5.72$; range 3.6 to 7.3) than taxonomic structure (mean $S_{TAX} = 1.64$; range 1.0 to 3.4), $p < 0.001$. In contrast, the *taxonomic-instructed group* showed no significant difference between ratings of S_{THEM} (mean 4.46; range 2.9 to 5.7) and S_{TAX} (mean 4.62; range 3.6 to 5.8), $p = 0.57$.

We checked whether the results of this initial ANOVA changed when each cluster was given equal weight (regardless of the number of items in it) when calculating S_{TAX} and S_{THEM} (see Appendix A for further details). This second ANOVA replicated the pattern of results of the first so our findings were not sensitive to the method of calculation used.

5.2.3.2. By items. We examined whether the results found across participants generalised across items by calculating the mean S_{THEM} and S_{TAX} ratings for each item for the 17 clusters (one for each of the taxonomically-instructed participants) which contained that item and, separately, for the 17 clusters (one for each of the thematically-instructed participants) which included that item. In Experiments 1 and 2 we found mainly taxonomic sorting for a substantial minority of items ($[S_{THEM} < S_{TAX}]$ for 45/140 and 35/150 items respectively). For the *taxonomic-instructed group* in Experiment 3 this improved to 61/150, and for this group in Experiment 4 it increased again, with $[S_{THEM} < S_{TAX}]$ for 96/150 items (mean −0.2, range +4.4 to −4.4) so, for the first time, the majority of items (nearly two-thirds) were sorted into mainly taxonomic clusters. Only two of the 150 items sorted by the *thematic-instructed group* had $[S_{THEM} < S_{TAX}]$ (mean +3.4, range +6.2 to −0.1).

5.2.3.3. Comparing sorting consistency across participants in Experiments 3 and 4. We added the data from Experiment 3 and repeated the by-participants ANOVA for S_{THEM} and S_{TAX} ratings as reported above. This ANOVA included a second, between-participants factor of random pile (allowed or not). As expected, ratings were, overall, higher in Experiment 4, when participants could use random piles (4.11), than in Experiment 3, when they could not (3.76), $F(1,55) = 5.99$, $p = 0.02$, partial $\eta^2 = 0.10$, whilst the pattern of results was similar across both experiments.

5.3. Discussion

In Experiment 4 most participants chose to produce a random pile of the items that they could not easily cluster. We found no difference between the number of items discarded into the random pile across the thematic-instructed and taxonomic-instructed groups (around 20% in both cases). Despite the extensive use of the random pile the results from Experiment 4 were similar to those of Experiment 3. Crucially, the taxonomic-instructed group still did not make clusters which were rated as having greater taxonomic than thematic structure. People did, though, sort more consistently when they could use the random pile, in Experiment 4, than when they could not, in Experiment 3. These results are consistent with the conclusion from Experiments 1, 2 and 3 that, although both thematic and taxonomic relations are encoded, there are more thematic than taxonomic relations stored in our long-term semantic knowledge and so thematic clustering dominates in large-scale sorting tasks.

5.3.1. Two further analyses investigating whether the nature of the semantic relations used to sort a given item predicts performance in other tasks

The results that we have reported so far suggest that there are more thematic than taxonomic relations available in our stored semantic knowledge to support basic-level sorting. However, if the two sets of stimuli used in Experiments 1–4 were not representative of the objects that we usually interact with these results might not reflect the relative number of thematic versus taxonomic relations available to us in our everyday life. This is the dog-kennel problem: in a sorting task, if the only available associate of kennel is dog then dog and kennel will have to be clustered together. If this occurred in our studies then our results may only tell us about sorting for the specific sets of items that we used, rather than informing us about the nature of the semantic relations stored in our semantic knowledge networks. Two approaches were taken to address this issue. The first approach examined whether the same type of relation was used to sort the subset of items that were included in both of the stimulus sets used for free-sorting in Experiments 1 and 2. The second approach investigated whether the type of semantic relation used to free-sort a given item in Experiments 1 or 2 predicted the type of single word associates generated to that item.

5.3.1.1. Comparing sorting consistency across items in experiments 1 and 2. First, we considered the 53 items which were similar across the sets of 140 and 150 items used in Experiments 1 and 2 respectively, see supplementary materials. For each item we compared the relative strength of thematic to taxonomic ratings $[S_{THEM}–S_{TAX}]$ across the two studies. There was a strong, positive correlation between $[S_{THEM}–S_{TAX}]$ in Experiment 1 and Experiment 2, $r(52) = +0.72$, $p < 0.001$. This suggests that similar semantic relations were used to sort these items across both studies even though the remaining two-thirds of items in the sets differed.

5.3.1.2. Generating word associations to items from Experiments 1 and 2 which were consistently sorted thematically or taxonomically. Second, we selected the 30 target items which were sorted most thematically (so where $[S_{THEM}–S_{TAX}]$ was greatest) and the 30 which were sorted most taxonomically (so where $[S_{THEM}–S_{TAX}]$ was least) in Experiments 1 and 2. We then investigated whether the type of semantic relation used when clustering each of these target items (thematic or taxonomic) predicted whether that item elicited predominantly thematic or taxonomic word associates. We looked up the free associations produced to the names of the 30 thematic and 30 taxonomic target items using the University of South Florida Word Association Norms (http://w3.usf.edu/FreeAssociation/Intro.html; Nelson, McEvoy, & Schreiber, 1998). Two of the 60 target items were duplicates and other items were absent from the database so 47 sets of word associates were obtained. An average of 151 US students responded to each of these items. Students wrote the first word that came to mind that was meaningfully related or strongly associated to the target item. Two of the current authors then

independently coded the relation of each of the 10 most frequently produced associates to a given target item as thematic, taxonomic or uncodeable.[4]

A chi-square test was conducted for each target item with the expected outcome taken as equal numbers of thematic and taxonomic responses. For the 28 items which were sorted taxonomically in Experiments 1 and 2, significantly more taxonomic than thematic associates were produced in 21 cases, the chi-square test was not significant in six further cases, and in just one case were significantly fewer taxonomic than thematic associates produced. Thus most items which were sorted taxonomically elicited taxonomic associates, with 21/22 significant differences in the predicted direction. For the 19 items which were sorted thematically in Experiments 1 and 2, significantly more thematic than taxonomic associates were produced for six cases, the chi-square test was not significant in seven further cases and there were six cases for which significantly fewer thematic than taxonomic associates were produced. Thus items which were sorted thematically were about as likely to elicit thematic as taxonomic associates, with 6/12 significant differences in the predicted direction.

Importantly, the nature of the semantic relations used to sort a given item (thematic or taxonomic) predicted the type of word associate generated to that item. This provides further evidence that our sorting results were not just due to the particular stimulus sets that we used. Overall, taxonomic relations were used more often than thematic relations to generate single word associates, reversing the thematic dominance which we observed for our large-scale sorting studies. In Experiments 1 and 2, the 30 items which were sorted most taxonomically had [S_{THEM}–S_{TAX}] ranging from −4.3 to just −1.0, whereas the 30 items which were sorted most thematically had [S_{THEM}–S_{TAX}] ranging from +5.7 to +4.5. Thus some of the taxonomic items were clustered only slightly more taxonomically than thematically whereas all of the thematic items were placed into strongly thematically related clusters. In contrast, in these word association analyses for those same items, 27/35 significant differences occurred because more taxonomic than thematic associates were produced, whereas only 8/35 significant differences occurred because more thematic than taxonomic associates were produced.

6. Experiment 5

A final study[5] was conducted because, in the word association analysis reported above, many of the single word associates generated to a target item could not be straightforwardly coded as revealing thematic or taxonomic relations between objects. For example, over half of the associates were excluded because they were not concrete objects. In Experiment 5 we collected new word associates which were generated to target items that were sorted either mainly thematically or mainly taxonomically in Experiments 1 and 2. Ambiguous responses were avoided by explicitly instructing people to generate names of concrete objects.

6.1. Method

6.1.1. Participants

Twenty adults volunteered to take part in the study.

6.1.2. Materials, design and procedure

The target items were the 10 most taxonomically sorted items ([S_{THEM}–S_{TAX}] was −4.3 to −1.8) and the 10 most thematically sorted items[6] ([S_{THEM}–S_{TAX}] was +5.7 to +4.4) in Experiments 1 and 2. These 20 items were presented to participants in one of two fixed orders (one the reverse of the other) which alternated thematically sorted and taxonomically sorted items. Each item was named aloud by the experimenter and participants generated a single word associate to it. Participants were told to name a concrete object that was the same kind of thing as the target item. Participants who produced non-acceptable responses (such as verbs or subordinate or superordinate labels) were reminded of their instructions and asked to generate another response.

6.2. Results

Two of the authors independently coded the relation of the 178 different word associates generated to the 20 target items as thematic (14%), taxonomic (61%) or uncodeable (25%). Disagreements were resolved by discussion. An independent samples *t*-test was conducted on the percentage of codeable responses classified as taxonomic. There was one between-items factor of semantic relations used to sort that item (taxonomic versus thematic). Significantly more taxonomic associates were generated to taxonomically sorted items than to thematically sorted items, $t(18) = 2.749$, $p < 0.03$. For the 10 items sorted taxonomically in Experiments 1 and 2, 99% of the word associates produced were taxonomic, whereas for the 10 items sorted thematically in Experiments 1 and 2, 64% of the word associates produced were taxonomic.

6.3. Discussion

The semantic relations used to sort a given item predicted whether thematic or taxonomic word associates were generated to that item: taxonomically-sorted items nearly exclusively elicited taxonomic word associates whereas thematic-sorted items generated around two-thirds taxonomic and one-third thematic word associates. Thus, overall, taxonomically related associates were produced more often than thematically related associates. We return to consider this result in the General Discussion. These results from Experiment 5, together with the two further analyses reported above, provide converging evidence that the results of the sorting studies reported here did not merely reflect the particular sets of stimuli used in Experiments 1–4. To summarise this evidence, first, there was a strong, positive correlation between the type of semantic relation (thematic versus taxonomic) used to free-sort the 53 items common to the two stimulus sets used in Experiments 1 and 2. Second, items which were consistently free-sorted taxonomically, in Experiments 1 and 2, were more likely to generate taxonomic word associates than items which were consistently free-sorted thematically. This was shown using pre-existing data from the University of South Florida Word Association Norms and, in Experiment 5, using data collected from a new group of 20 participants in a word association task.

7. General discussion

We used a large-scale, sorting task to try to gain insights into the nature of the relations between objects that are permanently stored in our semantic knowledge. This sorting task has similarities with many everyday tasks. We often need to decide how to cluster together large sets of objects from different basic level categories to allow ourselves or others to (re-)find items, for instance when we tidy our garage, put our shopping away or pack up to move house. These studies are the first that we know of to investigate how such unconstrained sorting tasks, using many objects from a wide variety of basic-level categories, are

[4] Disagreements between the two coders were resolved by discussion. Only associates that referred to concrete objects (not verbs, adverbs or adjectives) were coded. Parts of the target object were coded as thematic associates (e.g., wheel for motorbike) and subordinate and superordinate labels were coded as taxonomic associates. Some relations seemed to be both thematic and taxonomic (e.g., envelope: letter) whilst the nature of other relations was ambiguous (bus: school - is the relation to school bus, so a type of bus, or to the theme of school?) or were otherwise difficult to code (such as life-cycle relations like butterfly: caterpillar, and possible phrase completions like toilet: paper). All such responses were deemed uncodeable. We then calculated the total number of participant responses that were coded as thematic and as taxonomic for each target item.

[5] We thank Greg Murphy for suggesting the production experiment conducted to address this issue (Experiment 5).

[6] In addition, the most thematically sorted item from Experiment 1, the crucifix, was replaced by another item, pen, because pilot testing revealed that some participants were unsure of the exact referent of this word.

influenced by the overall structure of our stored, conceptual knowledge (although quite large free-sorting tasks have been used for objects from a single domain, for example Boster & Johnson, 1989, for 43 fish; Medin et al., 1997, for 48 trees; Medin et al., 2006, for 44 fish; and Ross & Murphy, 1999, for 45 food items). To our knowledge, our study involves the largest sets of items used in a free sorting study, and it provides a unique window into how people sort items from multiple domains.

Unlike typical small-scale sorting and matching tasks, the objects used here were not pre-selected to be related in a certain way to each other and there were few task constraints. Nevertheless, several lines of evidence indicate that people behaved similarly to each other rather than idiosyncratically, suggesting that this large-scale sorting task taps fundamental aspects of how adults represent semantic knowledge. First, in Experiment 1 there were strong correlations across the word group and the picture group for subjective ratings of S_{THEM} and S_{TAX} in the by items analysis. Second, the 53 items common to Experiments 1 and 2 showed high correlations for [S_{THEM}–S_{TAX}]. Third, in Experiments 1, 2 and 3 our novel, objective measure of Φ_{INTRA} showed that people sorted consistently with each other (see Table 1). People agreed with each other about how to sort the objects about as much as people agreed about the categorization of a large set of Munsell colour chips (Haslam et al., 2007). Although in Experiments 1 and 2 people were not told how to sort they were sensitive to much the same semantic relations such that they clustered objects in similar ways. This large-scale sorting task thus appears to tap stable, stored semantic knowledge of the world that is shared across people. Since it assesses how multiple items are clustered together it complements tasks such as matching and free association (e.g., Kiss, Armstrong, Milroy, & Piper, 1973; Nelson et al., 1998) which only indicate the strength of pairwise relations between items. In addition, our novel measures of Φ_{INTRA} and Φ_{INTER} showed that people were not sorting idiosyncratically in our studies. These measures should provide a useful means for objectively assessing sorting consistency in future research.

As outlined in the Introduction, at a fine scale of analysis, the semantic relations used in, for example, small-scale free-sorting tasks are both flexible and sensitive to the set of stimuli provided (Murphy, 2001), with people readily using both taxonomic and thematic relations to support categorization. We were, though, instead interested in the overall nature of the information stored in our semantic knowledge networks. We found that the use of thematic relations dominated sorting performance. When people sorted freely, in Experiments 1 and 2, they consistently clustered more thematically than taxonomically (80/81 participants; see also Fig. 1), across different presentation formats (words and line drawings and coloured pictures), across different sets of objects (Experiment 1 and 2), and even when people were primed to sort taxonomically (Experiment 2). Many free-sorted clusters were rated as having strong thematic but weak taxonomic structure (31%) whereas few were rated as having weak thematic but strong taxonomic structure (4%; see Fig. 2). Finally, even when explicitly instructed to do so, people failed to make mainly taxonomic clusters (for the taxonomic-instructed group in Experiment 3 and 4; see Fig. 1). Thus, for the objects tested here, there seemed to be more thematic than taxonomic relations stored in semantic knowledge and available for people to use to sort. Why? In general, there may be a wider range of types of thematic relations since these are less tightly defined than taxonomic relations. More specifically, there may be relatively few clear-cut cases of superordinate taxonomic categories, and these may principally comprise living things such as plants and animals. Lawson and Jolicoeur (2003, see also De Deyne et al., 2016) argued that many superordinate taxonomic categories of manmade objects are poorly specified (e.g., office equipment, household utensils, toys, games and electrical equipment). It remains for future research to determine whether certain characteristics, such as belonging to the broad category of living things, predicts whether an item is consistently sorted taxonomically. Approaches such as computational modelling using large sets of pre-existing data may help to address this issue.

The results from a word free association task (Experiment 5) indicated that, unlike the mainly thematic free-sorting of large sets of items (in Experiments 1 and 2), the generation of single word associates to concrete object categories may be mainly taxonomic. Thus the task (free sorting versus free association) may determine whether thematic versus taxonomic relations are accessed from semantic memory. Other researchers have reported differences in results based on word association data compared to other measures of semantic knowledge using a range of methodologies (e.g., Arias-Trejo & Plunkett, 2013; De Deyne et al., 2016; Greene, Baldassano, Esteva, Beck, & Fei-Fei, 2016; Koriat & Melkman, 1981; Steyvers & Tenenbaum, 2005). Taxonomic relations may be more salient or preferred, and so they may often be used to generate single word associates. However, if taxonomic relations are also less common than thematic relations this may make it difficult to use them to guide clustering of sets of items in a sorting task (see also De Deyne et al., 2016). It would be fruitful to investigate this directly by requiring different groups to free-sort the same set of items into small versus large clusters. If taxonomic relations are strong, but relatively few in number, then taxonomic sorting should be greater for the group told to make small clusters.

Importantly, despite this difference (thematic dominance for large-scale, free-sorting tasks versus taxonomic dominance for free association tasks), the results from free-sorting, in Experiments 1 and 2, predicted the type of words produced by free association in Experiment 5 (see also the analysis based on the University of South Florida Word Association Norms). Items which were sorted mainly taxonomically generated more taxonomic word associates than items sorted mainly thematically. This is evidence against the claim that the sets of stimuli used here did not reflect the distribution of objects that we interact with in everyday life. It will, though, be important to use different techniques and other stimulus sets to try to replicate our finding that the free-sorting of large sets of basic-level objects mainly produces thematically organised clusters. For example, object sets could be generated from word or free association databases (e.g., Miller, Beckwith, Fellbaum, Gross, & Miller, 1990; Nelson et al., 1998).

We propose that the dominance of thematic sorting provides evidence that, overall, there are many more thematic than taxonomic relations stored in our semantic knowledge networks. Note that this claim is independent of the participant's strategic *choice* to use taxonomic versus thematic relations when the number and strength of both types of relation are about equal. For example, if two similarly strong taxonomic and thematic alternatives are provided in a matching task then a given participant may overwhelmingly choose to match taxonomically. However, that same participant may, given a large-scale, sorting task, overwhelmingly sort thematically, if fewer taxonomic relations are *available*. This is because coherent clusters of several items cannot be based just on pairwise associations between items, however strong these may be.

We think that it is also important to highlight the difficulty in determining the nature of a given relation between objects. Researchers have sometimes assumed that there is a clear distinction between thematic and taxonomic categorization (Gentner & Brem, 1999; though see Estes et al., 2011). Furthermore, experimenters have normally tried to use stimuli which are related purely thematically or purely taxonomically. Together these two points may have masked the extent to which thematic and taxonomic relations may normally co-exist (Arias-Trejo & Plunkett, 2013; Jackson & Bolger, 2014). Our approach highlights the importance of such concerns and it permitted the use of thematic and taxonomic relations to be estimated independently. Estes et al. (2011); see also Lin & Murphy, 2001; Wisniewski & Bassok, 1999) tried to cleanly distinguish taxonomic from thematic relations. They argued that thematically related items play complementary roles in a scenario or an event, so cow + milk are thematically related because a cow produces milk. These relations include temporal, spatial, causal, functional, possessive and productive relations. However, on this basis many objects seem to be both taxonomically and thematically related. For example, the pairs of cup + teapot; chair + desk;

Table A1
Example contingency table.

		Participant 2		
		1	2	3
Participant 1	1	4 (a,b,c,d)	0	0
	2	0	3 (e,f,g)	1 (h)

avocado + lemon; sheep + sheepdog; and knife + fork have many common perceptual and functional properties but they also play complementary roles in events. One way to investigate this issue could be to try to measure different aspects of the psychological similarity between items in order to determine whether this predicts the types of relations that exist between objects.

Notwithstanding these multiple, competing or cross-cutting ways to categorize items (Estes et al., 2011; Ross & Murphy, 1999; Wisniewski & Bassok, 1999), in the present large-scale, free-sorting task the most common type of cluster produced had strong thematic structure together with weak taxonomic structure (31% of the total; see Fig. 2). Here, people clustered together items which shared little in common either perceptually or functionally (e.g., pushchair + child's dummy + teddy-bear; or bone + kennel + dog-lead). Taxonomic relations were also used, particularly early in a sort (see Experiment 2), and some items were mainly sorted taxonomically. However, pure taxonomic clusters were rarely produced (4%, see Fig. 2). Also, when people were explicitly told to sort using taxonomic relations they failed to sort exclusively taxonomically, whereas they were quite successful at sorting purely thematically when instructed to do so (see Experiments 3 and 4). It remains to be seen whether this balance between the proportion of thematic and taxonomic relations which can be accessed from our semantic knowledge differs across other tasks and other groups of people (Koriat & Melkman, 1981; Medin & Atran, 2004; Wolff, Medin, & Pankratz, 1999). Our results show that, although items may mainly be sorted into thematically organised clusters (see Experiments 1 and 2), the single word associates generated to those same items may mainly be related taxonomically (see Experiment 5). This suggests that, even for the same items, the nature of the semantic relations used to cluster multiple items together (mainly thematic) may differ from those used to generate single associates (mainly taxonomic). Further research will be needed to address why this occurs. For now we tentatively conclude that most of the information stored in our semantic knowledge is thematic rather than taxonomic. Consistent with this, De Deyne et al. (2016) recently demonstrated that both text-based corpora and word association data could be used to generate semantic networks and that these networks organised meaning thematically at the global (and at the local) scale of analysis. These thematically organised networks nevertheless included some taxonomic structure. Their results, like ours, suggest that our long-term, stored semantic networks are dominated by thematic information (storing the co-occurrence of concepts in space or time), although they do also store basic-level, taxonomic information (the perceptual and functional similarity of concepts). Such semantic networks can then support our ability to rapidly and flexibly use both thematic and taxonomic relations when we categorize.

Acknowledgements

We would like to thank Sarah Powell, Joanne Roberts, Hannah Quinn and Holly Quinlan for helping to test participants.

Appendix A. Dependent measures

A.1. Subjective ratings of the thematic and taxonomic structure of clusters (S_{THEM} and S_{TAX})

Based on the objects included in the cluster and the cluster name provided by the participant, five psychologists rated all 3603 clusters that were produced in the first four studies reported here and which were produced in a further, unpublished study which tested 31 participants on a replication of the control condition in Experiment 2 but with speeded sorting. In Experiment 4 ratings were only collected for the sorted items; no S_{THEM} or S_{TAX} ratings were obtained for the random piles. The raters were four post-graduate psychology students and the first author. The students were paid for their participation, naive as to the purposes of the study and were not involved in categorization research. All raters were blind to all experimental manipulations (e.g., which condition of which study a given cluster belonged to and which clusters were produced by a given participant) except that, because Experiment 1 used a different stimulus set, the first author knew which clusters came from that study.

The clusters were ordered so that similar and identical clusters were rated successively. The order was produced by sorting the clusters alphabetically by the labels given to the clusters and then by the first, second and third items included in the cluster. For example, the three clusters, labelled "DIY" and comprising [paintbrush + plug + lightbulb + screw]; [paintbrush + screw + broomstick]; [paintbrush + screw + bucket], were rated successively. One cluster order was given to two of the naive raters and the reverse order to the remaining two naive raters and to the first author.

The raters were instructed as a group but rated independently. Each rater gave every cluster two ratings using a scale from 0 (no coherent structure; a random set of items) to 9 (maximally coherent structure). One rating was for thematic structure and the other was for taxonomic structure. Raters were told that thematic clusters include things that belong together because they are found in the same place or the same event and/or are used together, regardless of whether they look the same or have the same properties or functions. They were told that thematically related objects often have complementary roles, such as teabag and teapot. They were given examples such as a thematic cluster name being "theatre" and including stage, costume, script, etc. Raters were instructed that taxonomic clusters include items that belong together because they are the same kind of thing and have a common set of properties. They were given examples such as a taxonomic cluster name being "jobs" and including teacher, postman, actor, etc. It was emphasised that a cluster could have both strong thematic and strong taxonomic structure or that it could have both weak thematic and weak taxonomic structure. For each cluster order one of the two naive raters rated taxonomic then thematic structure and the other naive rater and the first author rated thematic then taxonomic structure. Each rater took around 15 h to complete this task. Due to coding omissions the allocation to a cluster was not recorded for five items sorted in Experiment 2 and one item in Experiment 3.

To check for inter-rater reliability, one-tailed Spearman's rho correlations were calculated for each of the ten possible pairwise comparisons of the five raters. All ten pairings for taxonomic ratings were significantly correlated ($p < 0.001$) with the average $r(2915) = +0.57$ (range +0.42 to +0.74). In addition all but two pairings for the thematic ratings were significantly correlated ($p < 0.001$ for the eight significant cases), with average $r(2915) = +0.31$ (range −0.02 to +0.60). There was no evidence that the first author produced ratings that differed from the naive raters and her ratings correlated more highly with the four naive raters (mean of +0.64 for taxonomic structure and +0.43 for thematic structure) than each of the naive raters did with the remaining three naive raters.

S_{TAX} was then computed for each participant tested by taking the weighted mean of the five raters' taxonomic ratings of each of the clusters that that participant had produced. Weighting was by the number of *items* sorted into the cluster such that equal importance was given to the sorting of each of the items (which meant that larger clusters had more influence on S_{TAX}). To investigate whether this method of calculating S_{TAX} influenced the results S_{TAX} was then recalculated except that each *cluster* was given equal weight (regardless of the number of items in it; this meant that all clusters had equal influence on S_{TAX}).

This checked whether the initial S_{TAX} estimates, which were influenced more by the largest clusters, produced different results. This was not found to be the case. As reported in the results sections, ANOVAs produced similar results for both methods used to calculate weightings. S_{THEM} was computed from the five raters' thematic ratings in the same manner as S_{TAX}.

A.2. Objective measures of intra-group and inter-group agreement (Φ_{INTRA} and Φ_{INTRA})

Cramér's phi (Φ_C, Cramér, 1946) indexes the extent to which two categorizations are consistent with each other. In tasks such as matching where participants allocate one of an experimenter-defined set of labels to each stimulus, the consistency of two participants' categorizations can be indexed by simply calculating the percentage of agreement. In a free-sorting task, however, the experimenter does not explicitly provide category labels, and participants often differ substantially in the number of categories they produce. As Wills and McLaren (1998) have previously argued (see also Haslam et al., 2007), a statistic that indexes the level of association (i.e., co-prediction) between two categorical variables is a more appropriate measure of consistency between participants in a free-sorting task.

Cramér's phi provides such a measure. Specifically, it indexes the *consistency* between two different categorizations of the same set of stimuli – in our case, two attempts at free-sorting a large set of everyday objects. Φ_C has the desirable characteristic of varying from 0 (i.e., no association) to 1 (maximum association), irrespective of the number of clusters each participant produces. In the special case where both participants are using the same number of clusters, Φ_C is similar in magnitude to percentage agreement. For example, two people, each using 20 clusters, sort 150 items. Each participant produces 10 clusters containing 7 items and 10 clusters containing 8 items. They agree on the categorization of 80% of those items. Cramér's phi in this particular example is approximately 0.82.

Calculation of Φ_C begins by construction of a standard c by r contingency table, where c and r are the number of clusters used in the two classifications that are being compared. For example, if participant 1 uses 2 clusters and participant 2 uses 3 clusters, then a 2 by 3 contingency table is constructed where each count in that table represents one object they have both classified. To extend this example further, imagine there are just 8 items, that participant 1 partitions those items as [a,b,c,d], [e,f,g,h] and that participant 2 partitions the same items as [a,b,c,d],[e,f,g],[h]. The following contingency table results:

The next step is to compute the chi-square statistic (without correction for continuity) from this table (see e.g. Howell, 1992, p. 147–148). In our example (Table A1), $\chi^2(2) = 8$. Φ_C is then calculated as:

$$\Phi_c = \sqrt{\frac{\chi^2}{N(k-1)}}$$

where k is defined as the smaller of r and c and N is the number of stimuli that both participants have classified. In our Table A1 example $\Phi_C = \text{sqrt}(8/8) = 1$.

As should be apparent from the above example, calculation of Φ_C, unlike calculation of percent agreement, is not limited to cases where the two participants are using the same number of clusters. The mechanics of the Φ_C measure mean that the participant with the larger number of clusters receives no direct penalty for their finer grain of classification. Φ_C drops below 1 where the classification of the participant with the larger number of clusters is not simply a subdivision of the clusters of the participant with the smaller number of clusters. For example, if participant 1 produced [a,b,c,d],[e,f,g,h] and participant 2 produced [a,b,c,e],[d,f,g],[h], $\Phi_C = 0.53$.

In the current study, we employed two different measures based on Φ_C:

- Φ_{INTRA} - the average level of consistency between participants in the *same* experimental group. It was calculated as the mean Φ_C across all within-group pairs of participants. We have used this measure previously in an examination of the free sorting of a large set of colours by normal adults and a semantic dementia patient (Haslam et al., 2007).
- Φ_{INTER} - the average level of consistency between participant pairs *across* two different experimental groups. This was calculated as the mean Φ_C of all cross-group pairs of participants. This measure is novel.

Both measures can be the subject of inferential statistics, but standard analytic methods (e.g., t-tests) are not appropriate. This is because both Φ_{INTRA} and Φ_{INTER} measures are calculated across *pairs* of participants, and thus the data points that make up the mean are not independent (e.g., note that the number of pairs of participants in Φ_{INTRA}, $N(N-1)/2$, exceeds the number of participants, N). In the current article, we employed Monte Carlo methods (1×10^5 iterations) to estimate the probability of a Type I error for each of the following hypotheses:

Hypothesis 1. Intra-group agreement (Φ_{INTRA}) is greater than expected from chance responding, where chance responding is defined as using the same number of clusters as the participants, but otherwise allocating items to clusters at random.

Hypothesis 2. Inter-group agreement (Φ_{INTER}) is lower than would be expected if group membership were random, under the constraint that sample size for each group is preserved. Where inter-group agreement is lower for the actual allocation of participants to groups than would be expected for a random allocation, the inference is that the two groups are clustering the items differently to each other. If this is not apparent, consider the following example. Group A all sort the stimuli in one way (categorization X), Group B all sort the stimuli in a different way (categorization Y). Inter-group agreement is thus very low – no member of Group A sorts in the same way as any member of Group B. Now randomly mix up these two groups. Each group will now have about half X-style categorizers and half Y-style categorizers. So, now about half of Group A members agree with about half of Group B members. Inter-group agreement is thus moderate – and higher than it was before the random re-organisation of the groups. Thus, if two groups categorize differently, their inter-group agreement should be lower than would be expected from a random re-organisation of the groups. Chance levels of both hypotheses are affected by the number of clusters used by each participant. This means that Monte Carlo simulations must be run separately for each hypothesis tested, which is somewhat complex and computationally intensive.

The statistical analyses in this paper provide a novel contribution. Φ_{INTER}, and the associated numerical methods to determine chance levels, were developed specifically for the current paper. Φ_{INTRA} has been used in two previous studies which were not on the topic of taxonomic and thematic categorization (Haslam et al., 2007; Wills & McLaren, 1998). The present studies demonstrate that the combination of Φ_{INTER} and Φ_{INTRA} provide a useful tool for objectively investigating clustering in the free sorting of large stimulus sets. We encourage their use in future research, and have provided open-access source code and other materials to support this (www.willslab.co.uk/phi/).

Appendix B. Supplementary materials

Supplementary data to this article can be found online at http://dx.doi.org/10.1016/j.actpsy.2016.11.001.

References

Arias-Trejo, N., & Plunkett, K. (2013). What's in a link: Associative and taxonomic priming effects in the infant lexicon. *Cognition*, *128*(2), 214–227.

Boster, J. S., & Johnson, J. C. (1989). Form or function: A comparison of expert and novice judgments of similarity among fish. *American Anthropologist*, *91*, 866–889.

Cramér, H. (1946). *Mathematical models of statistics*. Princeton, NJ: Princeton University Press.

De Deyne, S., Verheyen, S., & Storms, G. (2016). Structure and organization of the mental lexicon: A network approach derived from syntactic dependency relations and word associations. *Towards a theoretical framework for analyzing complex linguistic networks* (pp. 47–79). Berlin Heidelberg: Springer.

Estes, Z., Golonka, S., & Jones, L. L. (2011). Thematic thinking: The apprehension and consequences of thematic relations. In B. Ross (Ed.), *The psychology of learning and motivation*, *Vol. 54*. (pp. 249–294). Burlington: Academic Press.

Gentner, D., & Brem, S. K. (1999). Is snow really like a shovel? Distinguishing similarity from thematic relatedness. *Proceedings of the Twenty-First Annual Meeting Of The Cognitive Science Society* (pp. 179–184).

Greene, M. R., Baldassano, C., Esteva, A., Beck, D. M., & Fei-Fei, L. (2016). Visual scenes are categorized by function. *Journal of Experimental Psychology: General*, *145*(1), 82.

Haslam, C., Wills, A. J., Haslam, S. A., Kay, J., Baron, R., & McNab, F. (2007). Does maintenance of colour categories rely on language? Evidence to the contrary from a case of semantic dementia. *Brain and Language*, *103*, 251–263.

Howell, D. C. (1992). *Statistical methods for psychology* (3rd ed.). Belmont, CA: Duxbury Press.

Jackson, A. F., & Bolger, D. J. (2014). Using a high-dimensional graph of semantic space to model relationships among words. *Frontiers in Psychology*, *5*.

Kiss, G., Armstrong, C., Milroy, R., & Piper, J. (1973). *An associative thesaurus of English and its computer analysis*. Edinburgh: Edinburgh University Press.

Koriat, A., & Melkman, R. (1981). Individual differences in memory organization as related to word-association, object-sorting, and word-matching styles. *British Journal of Psychology*, *72*(1), 1–18.

Landauer, T. K., McNamara, D. S., Dennis, S., & Kintsch, W. (2013). *Handbook of latent semantic analysis*. Psychology Press.

Lawson, R., & Jolicoeur, P. (2003). Recognition thresholds for plane-rotated pictures of familiar objects. *Acta Psychologica*, *112*, 17–41.

Lin, E. L., & Murphy, G. L. (2001). Thematic relations in adults' concepts. *Journal of Experimental Psychology: General*, *130*, 3–28.

Lopez, A., Atran, S., Coley, J. D., Medin, D. L., & Smith, E. E. (1997). The tree of life: Universal and cultural features of folk-biological taxonomies and inductions. *Cognitive Psychology*, *32*(3), 251–295.

Maguire, M. J., Brier, M. R., & Ferree, T. C. (2010). EEG theta and alpha responses reveal qualitative differences in processing taxonomic versus thematic semantic relationships. *Brain and Language*, *114*, 16–25.

Medin, D. L., & Atran, S. (2004). The native mind: Biological categorization and reasoning in development and across cultures. *Psychological Review*, *111*, 960–983.

Medin, D. L., Wattenmaker, W. D., & Hampson, S. E. (1987). Family resemblance, conceptual cohesiveness, and category construction. *Cognitive Psychology*, *19*(2), 242–279.

Medin, D. L., Lynch, E. B., Coley, J. D., & Atran, S. (1997). Categorization and reasoning among tree experts: Do all roads lead to Rome? *Cognitive Psychology*, *32*, 49–96.

Medin, D. L., Ross, N., Atran, S., Burnett, R., & Blok, S. (2002). Categorization and reasoning in relation to culture and expertise. In B. Ross (Ed.), *Psychology of learning and motivation*, *Vol. 41*.New York: Academic Press.

Medin, D. L., Ross, N. O., Atran, S., Cox, D., Coley, J., Proffitt, J. B., & Blok, S. (2006). Folk biology of freshwater fish. *Cognition*, *99*, 237–273.

Miller, G. A., Beckwith, R., Fellbaum, C. D., Gross, D., & Miller, K. (1990). WordNet: An online lexical database. *International Journal of Lexicography*, *3*, 235–244.

Milton, F., & Wills, A. J. (2009). Long-term persistence of sort strategy in free classification. *Acta Psychologica*, *130*, 161–167.

Mirman, D., & Graziano, K. M. (2012). Individual differences in the strength of taxonomic versus thematic relations. *Journal of Experimental Psychology: General*, *141*, 601–609.

Murphy, G. L. (2001). Causes of taxonomic sorting by adults: A test of the thematic-to-taxonomic shift. *Psychonomic Bulletin & Review*, *8*, 834–839.

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

Nelson, D. L., McEvoy, C. L., & Schreiber, T. A. (1998). The University of South Florida word association, rhyme, and word fragment norms. http://www.usf.edu/FreeAssociation/

Nguyen, S. P., & Murphy, G. L. (2003). An apple is more than just a fruit: Cross-classification in children's concepts. *Child Development*, *74*, 1783–1806.

Olver, R. R., & Homsby, J. R. (1966). On equivalence. In J. S. Bruner, R. R. Olver, & P. M. Greenfield (Eds.), *Studies in cognitive growth* (pp. 68–85). New York: Wiley.

Rosch, E., Mervis, C. B., Gray, W., Johnson, D., & Boyes-Braem, P. (1976). Basic objects in natural categories. *Cognitive Psychology*, *8*, 382–439.

Ross, B. H., & Murphy, G. L. (1999). Food for thought: Cross-classification and category organization in a complex real-world domain. *Cognitive Psychology*, *38*, 495–553.

Saalbach, H., & Imai, M. (2007). Scope of linguistic influence: Does a classifier system alter object concepts? *Journal of Experimental Psychology: General*, *136*, 485–501.

Shafto, P., Kemp, C., Mansinghka, V., & Tenenbaum, J. B. (2011). A probabilistic model of cross-categorization. *Cognition*, *120*, 1–25.

Simmons, S., & Estes, Z. (2008). Individual differences in the perception of similarity and difference. *Cognition*, *108*, 781–795.

Smiley, S. S., & Brown, A. L. (1979). Conceptual preference for thematic or taxonomic relations: A nonmonotonic age trend from preschool to old age. *Journal of Experimental Child Psychology*, *28*, 249–257.

Snodgrass, J. G., & Vanderwart, M. (1980). A standardized set of 260 pictures: Norms for name agreement, image agreement, familiarity, and visual complexity. *Journal of Experimental Psychology Human Learning and Performance*, *6*, 174–215.

Steyvers, M., & Tenenbaum, J. B. (2005). The large-scale structure of semantic networks: Statistical analyses and a model of semantic growth. *Cognitive Science*, *29*(1), 41–78.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., & Rabinovich, A. (2014). *Going deeper with convolutions. arXiv preprint arXiv:1409.4842*.

Tare, M., & Gelman, S. A. (2010). Determining that a label is kind-referring: Factors that influence children's and adults' novel word extensions. *Journal of Child Language*, *37*, 1007–1026.

Wattenmaker, W. D. (1995). Knowledge structures and linear separability: Integrating information in object and social categorization. *Cognitive Psychology*, *28*, 274–328.

Wills, A. J., & McLaren, I. P. L. (1998). Perceptual learning and free classification. *Quarterly Journal of Experimental Psychology*, *51B*, 235–270.

Wisniewski, E. J., & Bassok, M. (1999). What makes a man similar to a tie? Stimulus compatibility with comparison and integration. *Cognitive Psychology*, *39*, 208–238.

Wolff, P., Medin, D. L., & Pankratz, C. (1999). Evolution and devolution of folk biological knowledge. *Cognition*, *73*, 177–204.