MEMORY, 2003, *11* (6), 533–548

The instantiation principle re-evaluated

Els De Wilde, Veerle Vanoverberghe, Gert Storms, and Paul De Boeck
University of Leuven, Belgium

Three experiments are presented in which different aspects concerning Heit and Barsalou's (1996) instantiation principle were investigated. Mean typicalities of subordinate categories within superordinates were predicted very accurately for all investigated concepts. Multiple instantiations were shown to yield somewhat better predictions than single instantiation. The instantiation principle also successfully predicted mean typicalities on a different level (i.e., in lower-level concepts). An alternative account of Heit and Barsalou's findings was also proven wrong. Finally, correspondence between empirically obtained and predicted standard deviations is argued to be dubious, because of several possible sources of bias in the observed and predicted values.

Since the growth of cognitive psychology in the 1970s, natural language concepts have been studied extensively. (For overviews of the literature see, for instance, Komatsu, 1992; Medin & Smith, 1984; Smith & Medin, 1981). One of the most important findings in these studies is that concepts are not defined according to a set of features that are singly necessary and jointly sufficient. Different models have been proposed over the years to explain mental representation of natural language concepts. These models can be divided into two classes: models primarily based on the intension of the concept and models primarily based on the extension of the concept. An example of the former is Hampton's (1979, 1993) feature-based prototype model, which assumes that a concept is represented by a set of characteristic features that are directly stored with the concept label. Extension-based models, on the other hand, assume that concepts are represented by means of exemplars (e.g., Medin & Schaffer, 1978; Nosofsky, 1986). Most evidence for exemplar models, however, comes from category-learning experiments in which participants learn new categories of artificial stimuli. Applications of exemplar models to natural language concepts are almost absent in the literature.

One remarkable exception is Heit and Barsalou's (1996) paper on the instantiation principle. The instantiation principle states that the representation of a category includes detailed information about its diverse range of instances. The principle was tested in a study in which Heit and Barsalou predicted typicalities of lower-level concepts (like, e.g., *mammals* or *beverages*) within a higher-level concept (*animals* and *food*, respectively) from the typicality of instances of the lower-level concept (like, e.g., *cat*, *dog*, *horse*, or *beer*, *coke*, *milk*). Across three studies, the authors reported that the model ''did an excellent job of predicting mean typicality judgments (...), and a good job of predicting standard deviations'' (Heit & Barsalou, 1996, p. 413).

In this paper, we want to address two questions concerning the instantiation principle, both of which have already been touched on briefly by Heit and Barsalou (1996). First, we want to evaluate whether activation of more than one instantiation for each category leads to better predictions than the Heit and Barsalou model,

Requests for reprints should be sent to Gert Storms, Psychology Department, University of Leuven, Tiensestraat 102, B-3000 Leuven, Belgium. Email: gert.storms@psy.kuleuven.ac.be

This research project was sponsored by Grant G.0353.99 from the Funds for Scientific Research—Flanders to Gert Storms and Paul De Boeck. We wish to thank Lawrence Barsalou and Evan Heit for discussion and helpful comments.

All the data described in this paper can be obtained from Gert Storms (in Excel files) upon simple request.

© 2003 Psychology Press Ltd
http://www.tandf.co.uk/journals/pp/09658211.html DOI:10.1080/09658210244000126

which assumes that only exactly one exemplar is instantiated. This question will be addressed in Experiments 1 and 2. Second, we want to evaluate (in Experiment 3) whether the instantiation principle can also be applied to predict typicalities within lower-level concepts. Due to the central importance of the Heit and Barsalou model in this paper, we will first review the main ideas of their instantiation model.

A SHORT REVIEW OF HEIT AND BARSALOU'S (1996) INSTANTIATION MODEL

Heit and Barsalou (1996) developed and tested an exemplar-based model of instantiation. According to the model, people judge typicality of a subordinate in a superordinate by performing three steps. First a *single* instantiation of the subordinate category is retrieved (e.g., *cow* is retrieved to instantiate *mammal*). Second, the instantiation's typicality in the superordinate is evaluated (e.g., the typicality of *cow* in *animal* is judged on a 9-point rating scale, ranging from 1 for very atypical to 9 for very typical). Let's assume that *cow* gets a score of 7 on the typicality scale. Finally, the instantiation's typicality is generalised to the subordinate (e.g., the typicality of *mammal* in *animal* is rated as 7). Thus, the only information the model uses to predict performance is a distribution of instantiations of the subordinate (i.e., the *first* instantiation that is activated given the category label) and the typicalities of these instantiations in the superordinate.

Described formally, the model amounts to

$$P(T(A) = j) = \sum_{i=l}^{n} P(A \textit{ is instantiated as } a_i)\, P(T(a_i) = j),$$

where $T(x)$ is the typicality of x within the superordinate category, A is a subordinate category, a_i is an instantiation of A, n is the number of different instantiations of A, and j is an integer on a predefined rating scale. Given this model, the distribution of typicality ratings for the subordinate category (e.g., *mammals*) within the superordinate category (e.g., *animals*) may be simulated using production frequencies of (first) instantiations of the subordinate category (e.g., *cow*) and typicalities of these instantiations in the superordinate category.

Heit and Barsalou (1996) reported three experiments. In the first experiment, they correlated, for seven subordinates of the superordinate category *animal*, mean empirically obtained typicalities with mean values of the simulated typicality distribution, and also the standard deviations of the empirically obtained and of the simulated distributions. The second experiment replicated these findings for nine subordinates of the superordinate category *food*. In the third experiment, the instantiation principle was applied to complex categories like *dangerous animals*. For the data from Experiment 3, the skewness coefficients of the empirically obtained and the simulated distributions were also correlated. The correlations for the distribution means were high (generally above .90) in all three experiments, but the correlations for the distribution standard deviations were a bit lower (between .60 and .90). For the skewness (Experiment 3) a correlation of .87 was found. Heit and Barsalou interpreted these results as support for the instantiation principle; that is, they claimed their findings provide evidence that subjects' typicality judgements for members of a category are influenced by detailed information about the member's instantiations.

EXPERIMENT 1

In the first experiment, we wanted to evaluate Heit and Barsalou's (1996) assumption that only one single instantiation is activated in judging typicalities of subordinates in superordinate categories. More specifically, we wanted to evaluate whether a model that incorporates information of more instantiated exemplars (not just the *first* instantiated exemplar) improves the prediction of the typicalities. Imagine, for example, evaluating the typicality of the subordinate category *fruit* within the superordinate category *food*. According to a multiple-instantiation model, subjects do not only use idiosyncratic knowledge about the first kind of *fruit* that comes to mind to assess the typicality of *fruit* (e.g., *apple*), but they add knowledge based on subsequently activated instances (e.g., *orange*, *pear*, *banana*, *grape*, *pineapple*, ...).

In their first two experiments, Heit and Barsalou correlated means and standard deviations of the empirically obtained typicality rating distribution and of a simulated distribution. The simulated distribution was based (1) on rated typicalities (in the superordinate category) of exemplars of the subordinate, and (2) on the

generation frequencies of these exemplars, taken from a task in which participants were instructed to name the *first* instantiation that came to mind given the category cue. The idea behind this procedure is that the exemplar-generation task (of *first* instantiations only) samples from the same distribution as the instantiation process that guides the typicality ratings of the subordinates. The assumption that only one instantiation is activated can be questioned on the ground of at least two arguments. First, most exemplar models in the category-learning tradition (e.g., Medin & Schaffer, 1978; Nosofsky, 1988) assume an activation process in which *all* known exemplars are retrieved, rather than just one. Second, within the context of natural language concepts, direct evidence for a multiple instantiation process is given by Storms, De Boeck, and Ruts (2000), who showed that, for superordinate-level concepts, the summed similarities of seven to ten frequently generated instantiations result in the best predictions of typicality ratings, response times, category naming frequencies, and exemplar-generation frequencies. They found that, for categories like *fruit* (or *sports*), almost all subjects generate the same familiar exemplars like *apple* and *orange* (or *basketball* and *football*) first, whereas other well-known exemplars that were shown to improve the prediction, are instantiated only when more than one exemplar has to be generated.

Heit and Barsalou (1996) already recognised the possibility of multiple instantiations. They pointed out that, if subjects do retrieve multiple instantiations for a category, then their judgements should be more stable than if they retrieve only a single instantiation, ''just as the standard error of a mean decreases as sample size increases'' (p. 443). However, it is also possible that the more instantiations the subjects retrieve, the more similar the standard deviations of typicality judgements of the subordinate categories become, as they all decrease.

The hypothesis of multiple instantiations can easily be compared with the Heit and Barsalou assumption of single exemplar instantiation by taking exemplar frequencies from a task in which participants are instructed to name *many* instances of the presented subordinate category *in the order that they come up in their mind*. Different typicality rating distributions can then be simulated: one distribution based only on the frequencies of the instances given as the first response, and other distributions also based on the exemplars (and the corresponding frequencies) instantiated by the participants after their first instantiation. The first distribution then simply leads to correlations that replicate Heit and Barsalou (1996), and these correlations can be compared with correlations based on the simulated distribution of the first response plus the subsequent responses. The latter distribution may contain a broader range of instantiated exemplars.

Method

Participants. A total of 20 participants (13 undergraduate psychology students, 3 graduate students, and 4 relatives of the second author) volunteered in the exemplar production task, and 20 first-year psychology students of the Leuven University participated in the typicality rating task.

Materials. The material used was the category used in Heit and Barsalou's (1996) second experiment. Thus, for the exemplar-production task, the category cues were nine subordinates of *food*: *beverage*, *dairy product*, *dessert*, *fish*, *fruit*, *meat*, *poultry*, *seasoning*, and *vegetable*. For the typicality rating task, the rated categories were the nine subordinates plus 223 unique instantiations that the exemplar-production group generated.

Procedure. As in Heit and Barsalou (1996), the participants in the exemplar-production task performed the experiment individually. Each subject was instructed to write down the first *nine* instances that came to mind for each category cue, in the order that they thought of them.

Participants in the typicality rating task were instructed to rate the nine subordinate categories and the 223 instances on typicality with respect to *food*, using a 10-point rating scale (from 1 to 10), for which higher numbers meant greater typicality. The list of items, consisting of instantiations as well as subordinate categories, was ordered alphabetically. For example, the participants were asked, ''How typical is *fruit* for the category *food*?'' and ''How typical is *apple* for the category *food*?''. As in Heit and Barsalou's study, the participants never rated the typicality of the instantiations (e.g., *apple*) in the subordinate categories (e.g., *fruit*). The participants performed the rating task at their own pace, normally taking about 15 to 20 minutes.

Results and discussion

As in Heit and Barsalou (1996), the nine predicted distributions, obtained by application of the instantiation model, were compared with the nine observed distributions of the typicality ratings. The observed distribution for a subordinate consisted of the ratings that the 20 participants in the typicality rating task made for it directly. For each subordinate, means and standard deviations of the 20 participants were computed. The predicted distribution resulted from a simulation of the instantiation process. Different sorts of simulation processes were run: a simulation based on the assumption of only one single instantiation per subordinate and several simulations based on the assumption of an increased number of instantiations.

In the *single instantiation simulation*, a subordinate was instantiated according to the frequencies with which the exemplars were produced *as the first generated instantiation* in the exemplar-production task. For instance, 16 out of 20 participants produced *apple* as the first instantiation of the subordinate *fruit*, 3 produced *banana* first, and 1 produced *mandarin* first. Consequently, *apple* was used 80% of the time, *banana* was used 15%, and *mandarin* was used 5% of the time, on this step to predict the typicality of *fruit*. Next, a typicality for the instantiation (e.g., for *apple*) was chosen randomly from the 20 typicality judgements (i.e., the set of judgements from the 20 participants in the rating task). Replicating the procedure of Heit and Barsalou, a total of 400 participants were simulated for each subordinate in 20 cycles of 20 simulated participants each, ensuring that the relative frequency of the different instantiations was maintained in the simulated data, and that all rating data entered into the simulations.[1] For each subordinate, means and standard deviations were computed based on the 400 simulated participants.

For *the multiple instantiation simulation*, four different simulations were run, corresponding to a number of instantiated exemplars (n) varying from 2 to 5. For each of these 36 simulated distributions (9 subordinate categories by 4 different values of n), 50,000 participants were simulated as follows. A first instantiation was generated according to the frequencies of the first activated instantiations in the generation task. For this instantiated exemplar, a typicality rating was randomly selected based on the frequencies of the ratings given for the corresponding exemplar in the typicality rating task. Next, a second instantiation was randomly selected based on the frequencies of the second-activated instantiations in the generation task, but with the condition that an instantiation already selected as first instantiation could not be re-sampled (i.e., the sampling was done without replacement). This procedure was repeated n times, with n varying from 2 to 5. Then, a mean was calculated over the n sampled instantiations. Finally, means and standard deviations were calculated for each of these four distributions. In this way, simulated data could be compared assuming that participants instantiate the subordinates with 2–5 different exemplars.

Prediction of the means. For the single-instantiation model, the correspondence between the observed and the predicted means based on the simulations was quite high: $r = .90$ ($p < .01$). The high correlation was not attributable to extreme values, as is illustrated by the high Spearman rank correlation for the same data: ($r = .70$, $p < .05$). These values are comparable to the values reported by Heit and Barsalou (1996) ($r = .89$ and $r = .79$, respectively) for the same superordinate and subordinate categories.

For the multiple-instantiation model, correlations were calculated between the means of the empirically obtained data and of the simulated data, based on the first n instantiations of the subordinates, with n varying from 2 to 5. Figure 1 shows the resulting correlations as a function of n. The correlation increases from .90 for the single instantiation assumption to .94 for three activated instantiations. This indicates that taking into account multiple instantiations resulted in somewhat better predictions, but the prediction did not improve after adding more than three instantiations. However, it is not clear how much importance should be attached on the exact increase of the correlation for the following reasons: (1) the correlations are based on only nine data points, and (2) the improvement is rather small, but given the high values, large improvements cannot be expected due to a ceiling effect. (The latter argument holds regardless of the number of data points on which the correlation was calculated.)

[1] The simulation exhaustively sampled the same 20 exemplars instantiated first by all 20 participants for the subordinate in the exemplar-production task. The simulation sampled the typicality ratings for a particular instantiation without replacement until the entire set of ratings had been sampled. Thus, all 20 ratings for a given instantiation were sampled once each.

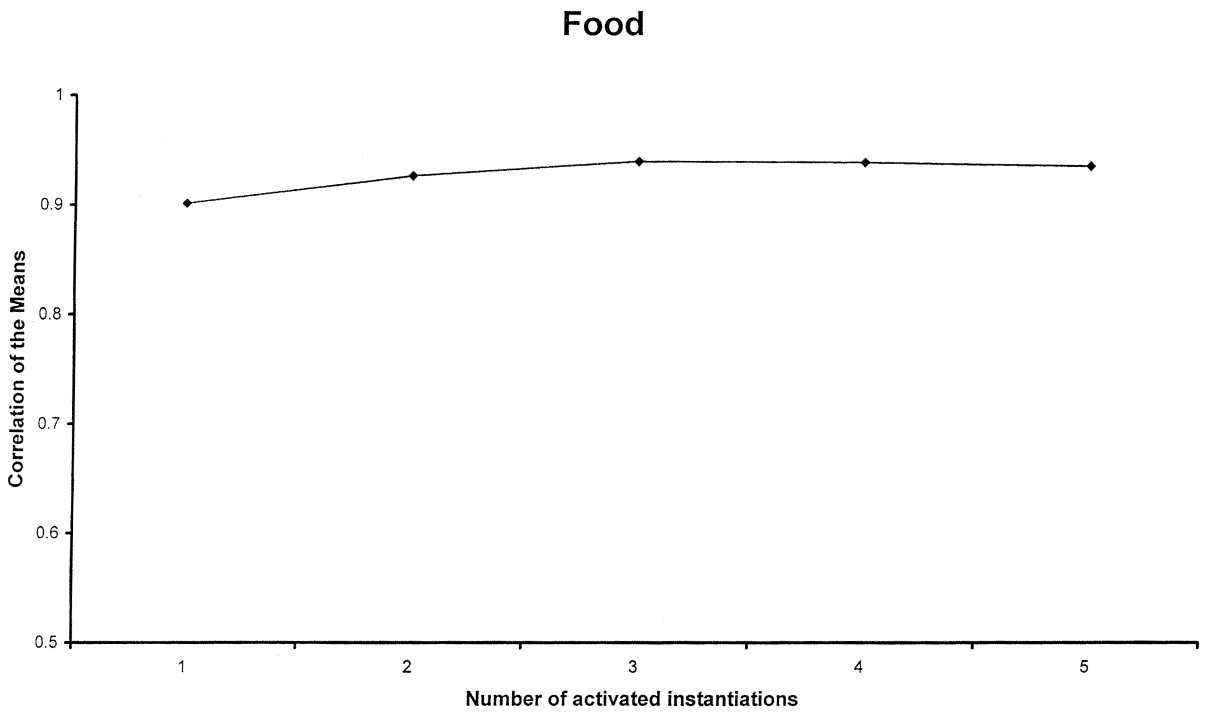


Figure 1. Correlations of observed and predicted typicalities as a function of the number of activated instantiations in Experiment 1.

More explicitly, with an estimated coefficient of .94 for the typicality ratings, correction for attenuation shows that the correlation between empirically obtained and predicted typicalities can never be larger than .97.[2]

An alternative explanation. The data reported earlier and the data of Heit and Barsalou (1996) were analysed using correlational methods and therefore do not allow causal interpretations. Moreover, it is possible that the correlations result from a representational process that is completely the opposite of what is assumed in the instantiation principle. The instantiation principle assumes that subjects base their typicality ratings of the subordinate categories on typicalities of instantiations of these subordinates. In other words, the instantiation principle assumes that the representation on which the typicalities are based is situated at least a level lower than the subordinate categories. However, it is possible that the typicalities of both the subordinates and the instantiations are based on the representations of the subordinate categories. In that case, subjects would reason as follows: ''I have to evaluate how typical a *pineapple* is for the category *food*. A *pineapple* is a *fruit*, and I judge *fruit* to be an 8 on the 10-point rating scale. Therefore, *pineapple* gets an 8 too.'' This reasoning may not seem very likely intuitively, but the correlations reported earlier and in Heit and Barsalou (1996) do not exclude this alternative interpretation.

A consequence of this alternative interpretation, however, is that there should be no significant differences between the mean typicality ratings of the different instantiations of a subordinate category. (Note that the reverse is not true: If there are no significant differences among the instantiations of a subordinate category, the instantiation principle might still apply.) Nine analyses of variance were done on the typicality ratings, one for each subordinate concept, to find out whether the mean typicalities of the instan-

[2] The correction for attenuation formula (Lord & Novick, 1968) states that the correlation between two variables X and Y, after correcting for unreliability, equals

$$\frac{r(X,Y)}{\sqrt{r_{XX}r_{YY}}}$$

where r_{XX} and r_{YY} indicate the reliability of the X and the Y variable, respectively. Given that correlations are always less than or equal to 1.0, one can deduce that the empirically obtained correlation is always less than or equal to $\sqrt{r_{XX}r_{YY}}$.

tiations differed significantly. Significant differences were found for all nine subordinate categories (all $p < .01$). These results are incompatible with the alternative interpretation.

Prediction of the standard deviations. For the single-instantiation model, the correlation between the observed and the predicted standard deviations was also high: $r = .84$ ($p < .01$). The Spearman rank correlation was $r = .93$ ($p < .01$). When calculating the correlation over the variances instead of the standard deviations, an even higher value was obtained ($r = .89$). These values are somewhat higher than the values reported by Heit and Barsalou for the same superordinate and subordinate categories ($r = .64$ and $r = .65$). The experiment thus succeeded in replicating Heit and Barsalou's (1996) findings.

For the multiple-instantiation model, similar analyses were done as for the means. Figure 2 shows the correlation for the variances as a function of the number of activated instantiations, varying from $n = 1$ to $n = 5$. The correlation decreases from .84 to .69. However, as can be expected from averages calculated over an increasing number of values sampled from a particular distribution, the standard deviations decreased in absolute size as the number of assumed instantiations increased from 1 to 5. The empirically obtained values, though, come from typicality ratings on a scale with a fixed range (from 1 to 10). It is quite likely that subjects in a rating task attempt to distribute their ratings as much as possible, thereby using the whole range of the scale (see, e.g., Jones, 1982; Storms, De Boeck, Hampton, & Van Mechelen, 1999). As a consequence, the decreasing correlations for increasing numbers of instantiated exemplars might be an artefact due to restriction of range.

Correlating different sources of variance. The results reported earlier do not unanimously support either the single instantiation, or the multiple instantiation hypothesis. On the one hand, means are best predicted by assuming two or three activated instantiations. On the other hand, standard deviations are better predicted by assuming only one activated instantiation, although this finding might be explained by restriction of range. Because of the lack of consistency in the results, the correlation between the observed and the predicted variances was investigated further by splitting up the predicted variance into variance (mean squares) due to inter-individual differences

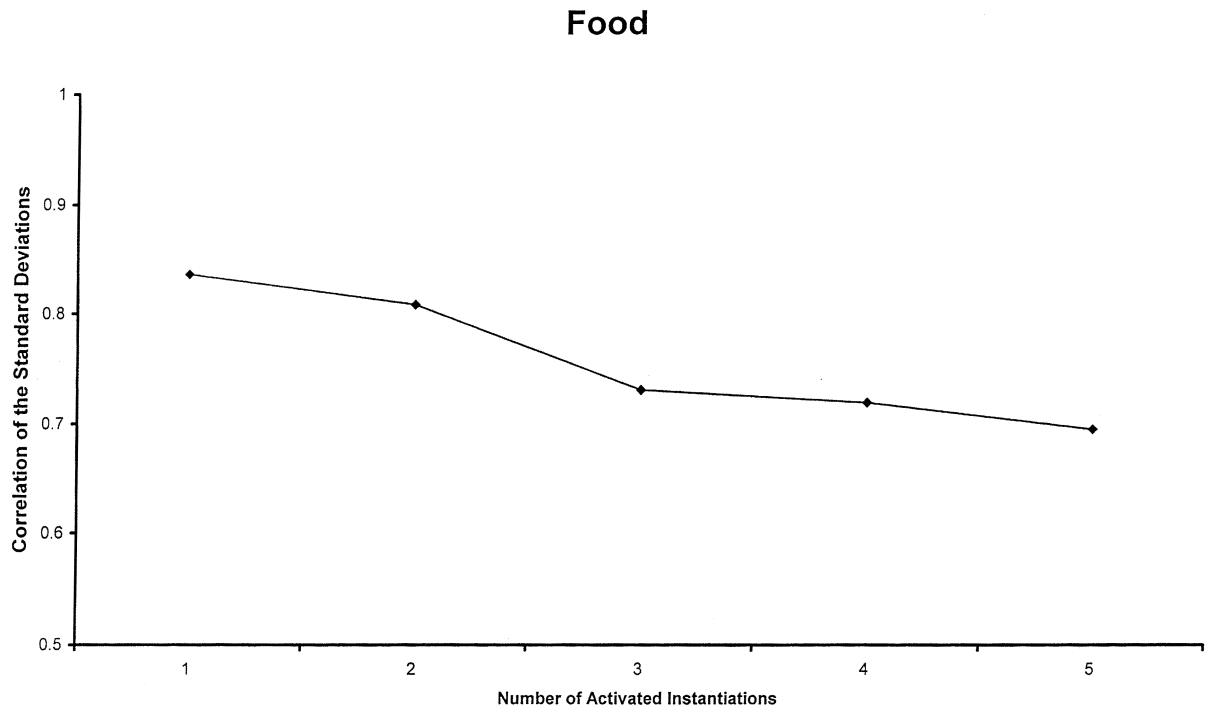


Figure 2. Correlations of observed and predicted standard deviations as a function of the number of activated instantiations in Experiment 1.

in the typicality ratings, and variance (mean squares) due to differences among instantiated exemplars. According to the instantiation principle, participants base their typicality ratings for the subordinates on the typicalities of the instantiated exemplars. Consequently, the variance in the typicality ratings of the subordinates contains two sources of variance: (1) variability due to different exemplars instantiated by the participants, which we will call *between-exemplar* variability, (for the category *animals* this means differences in typicality ratings between, for instance, *cows*, *dogs*, etc.), and (2) variability due to inter-individual differences in the typicality evaluation of the activated instantiations, which we will call *within-exemplar* variability (e.g., different typicalities given by different subjects for *dogs*). However, it is impossible to tear these two sources of variability apart in the typicalities of the subordinate. The simulated data of the single-instantiation study contain the same two sources of variance, but for these data, the two sources *can* be separated, using a simple analysis of variance, with the different instantiated exemplars as the levels of the independent variable. The resulting model mean squares correspond to the between-exemplar variance and the error mean squares correspond to the within-exemplar variance. The variance in the observed typicalities correlated .04 (not significant) and .90 ($p < .001$) with the between- and within-exemplar mean squares, respectively. In the simulated data the total mean squares correlates highly with the mean squares within exemplars ($r = .918$), but it also correlates .344 with the mean squares between exemplars. This difference between the empirical and the simulated data indicates that the correlation for the standard deviations found by Heit and Barsalou (1996), which was replicated in our data, was virtually completely based on a correspondence between the variances of the observed typicalities and the lack of unanimity among the participants in their typicality judgements. Some of the subordinate categories yield large inter-individual differences in their typicality ratings and in the typicality ratings of their instantiations, while other subordinate categories do not. This suggests that the correlation between the variances just reflects high inter-subject instability over participants in the rating task.

The correlation due to the within-exemplar variability is compatible with Heit and Barsalou's (1996) instantiation model, but is not very characteristic of the model. For instance, it is also compatible with the alternative interpretation presented earlier. Note also that the absence of a significant correlation of the between-exemplar variance and the empirically obtained variance cannot be attributed to the instantiations being too similar: There *are* significant differences in the instantiations activated by the subordinate concept labels, as was shown by the analyses of variance reported earlier. Still, these differences are *not* reflected in the correlations of the standard deviations, which only reflect within-exemplar variance.

In conclusion, the results of the first experiment replicated the findings of Heit and Barsalou (1996): high correlations were found between the means and the standard deviations of empirically obtained and simulated data. Looking at the means, the predictions of a multiple-instantiation model are somewhat better than those of the single-instantiation model, but adding more than three exemplars did not improve the prediction. For the standard deviations completely opposite results were obtained, which seemed to support the single-instantiation model. However, this finding might be attributable to restriction of range. Moreover, additional analyses showed that the correlation for these standard deviations was completely determined by within-exemplar variability and not at all by between-exemplar variability, a finding that is not very characteristic for the instantiation principle.

The small number of data points on which the correlations are calculated forces us to be careful in drawing strong conclusions. Experiment 2 was therefore set up to study a larger sample of subordinate categories.

EXPERIMENT 2

The main purpose of this experiment was to compare the single-instantiation assumption with the multiple-instantiation model in a concept with a larger set of subordinate categories. In Experiment 1, the predictive correlations for the single- and multiple-instantiation models were based on only nine subordinate categories. Although two of the three experiments described in Heit and Barsalou (1996) were also based on only nine or fewer data points, a category with a larger set of subordinate categories was desirable in order to obtain more reliable correlations.

As a solution to the small number of data points, Heit and Barsalou's Experiment 3 inves-

tigated complex categories (like *dangerous mammals* and *small birds*) as subordinate categories of either a simple superordinate (i.e., *animals*), or complex superordinates (i.e., *dangerous animals* and *small animals*). Their procedure of studying complex categories allowed them to use 21 subordinate categories in three different (simple or complex) superordinates.

Because of the complex processes involved in combining concepts into a conceptual combination (Wisniewski, 1997), a more suitable way to increase the set of subordinate categories was simply to look for a superordinate category with a large number of subordinates of which multiple instantiations are lexicalised and can be rather easily produced. Therefore, the category *clothing*, with 20 subordinate categories, was chosen.

Method

Participants. A total of 20 participants (all graduate students) volunteered in the exemplar-production task and 20 final year high-school students participated as volunteers in the typicality rating task.

Materials. The material used was the category *clothing*. For the exemplar production task, the category cues were 20 subordinates of *clothing*: *accessories*, *baby clothing*, *beach clothing*, *bedroom clothing*, *coats*, *evening dresses*, *free time clothing*, *headgear*, *rain clothing*, *shoes*, *skirts*, *sports clothing*, *stockings*, *sweaters*, *traditional clothing*, *trousers*, *underwear*, *uniforms*, *winter clothing* and *working clothing*. (The study was done using Dutch subordinate category labels. Some of these labels may not cover exactly the same extensions as the best English translations given here.) For the typicality rating task, the rated categories were the 20 subordinates plus 446 unique instantiations that the exemplar-production group produced.

Procedure. As in the first experiment, the participants in the exemplar-production task performed the experiment individually. Each subject was instructed to write down the first *seven* instances that came to mind for each category cue, in the order that they thought of them. They were strongly encouraged to complete a list of seven exemplars for every subordinate category, even if it was hard for some of the subordinates. (It was anticipated that for some of the subordinates, generating nine different exemplars was extremely hard or impossible. Therefore, we asked them to give seven exemplars, not nine as in Experiment 1. Not all participants succeeded in completing the list of seven exemplars for all subordinates, but for every list at least five exemplars were given.)

Similar to the procedure in Experiment 1, the participants in the typicality rating task were instructed to rate the 20 subordinate categories and the 446 instances on typicality with respect to *clothing*, using a 10-point rating scale, for which higher numbers meant greater typicality. The participants performed the rating task at their own pace, normally taking about 25 to 30 minutes.

Results and discussion

As in Experiment 1, two sorts of simulation processes were run: a simulation based on the assumption of only one single instantiation per subordinate, and a simulation based on the assumption of multiple instantiations. The simulation procedures were exactly the same as in Experiment 1.

Prediction of the means. The correspondence between the observed and the predicted means based on single-instantiation simulations was again very high: $r = .92$ ($p < .01$). Spearman's rank correlation was $r = .93$ ($p < .01$). For the multiple-instantiation predictions, Figure 3 shows the correlations between the empirically obtained means and the means of the simulated data as a fuction of the number of assumed instantiations. The correlation increased from .923 for $n = 1$ to .935 for $n = 5$. Thus, again the correlation was somewhat better when assuming multiple instantiations instead of single instantiation, but the difference was small. (Given the reliability estimates for the typicality ratings, the correction for attenuation formula showed that the correlation cannot be higher than .965.)

The alternative explanation of a representation of the instantiations at the subordinate level was evaluated again by performing analyses of variance on the ratings for each of the 20 subordinate categories. The results yielded significant differences ($p < .01$) for all subordinates. These data again plead against the alternative explanation.

Prediction of the standard deviations. The correlation between the observed and the

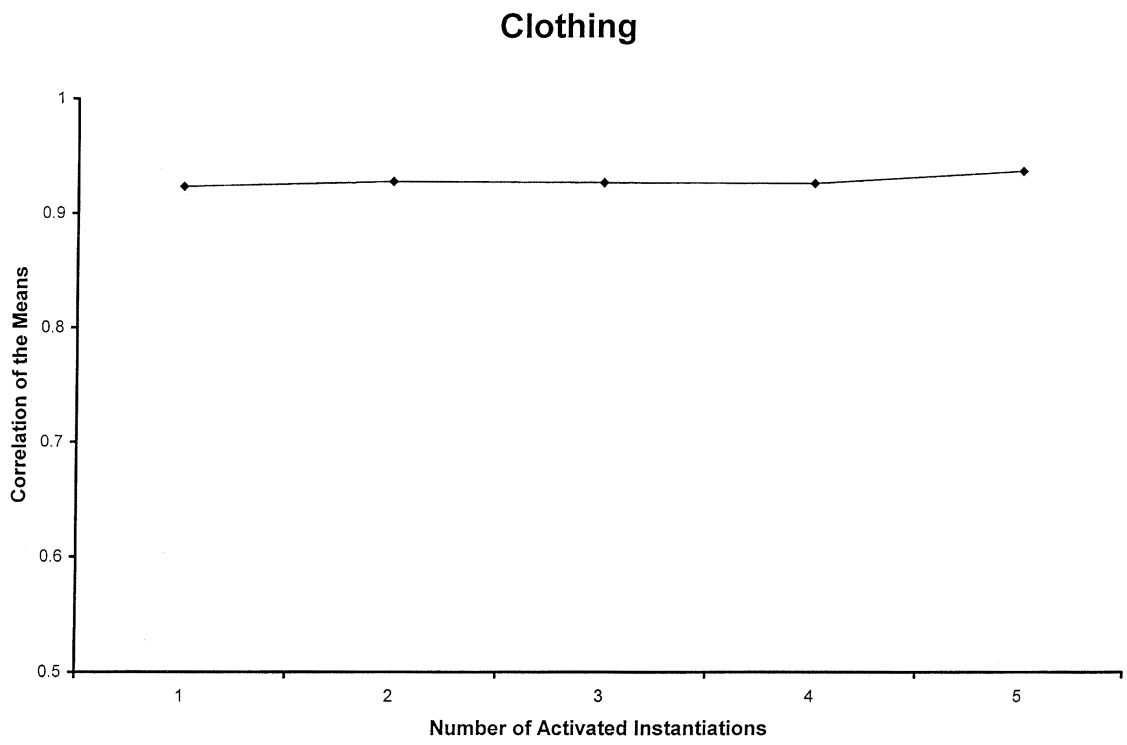


Figure 3. Correlations of observed and predicted typicalities as a function of the number of activated instantiations in Experiment 2.

predicted standard deviations was also high (r = .61, $p < .01$), with a rank correlation of $r = .70$ ($p <$ .01). The results of the multiple-instantiation model are shown in Figure 4. As in Experiment 1, the correlations decrease monotonically from .61 to .36. As in Experiment 1, the standard deviations decreased on absolute size as the number of assumed instantiations increased from 1 to 5. Therefore, the decreasing correlations for increasing numbers of instantiated exemplars might again be an artefact due to restriction of range.

As in Experiment 1, the between-exemplar variance of the single instantiation simulated data was separated from the within-exemplar variation. The variances of the empirically obtained rating data correlated .16 with the former (which is not significant) and .42 ($p < .05$) with the latter. Thus, despite the finding that there *are* indeed significant differences between the instantiations of each of the subordinates, the variance in the rated typicalities for the subordinates fails to correlate with the between-instantiation variation. Moreover, the total mean squares of the simulated data correlated .716 with the mean squares between exemplars (of these same simulated data) and .690 with the mean squares within exemplars (of the simulated data). This finding, which is similar to what was found in Experiment 1, is somewhat problematic for the instantiation principle.

In conclusion, the results of Experiment 2 are almost identical to the findings of Experiment 1. High correlations were found between the means and the standard deviations of empirically obtained and simulated data. The means were better predicted by a multiple-instantiation model than by the single-instantiation model. Despite the fact that the difference was small, we think it should be taken seriously because (1) the correlations of Experiment 2 were based on a large number of data points, and (2) due to a ceiling effect larger improvements could not be expected. For the standard deviations, opposite results were again obtained: single instantiation resulted in the highest correlation, but once more these findings might be attributable to restriction of range for the multiple instantiation data. The correlation for the standard deviation was again shown to be determined almost completely by within-exemplar variability and not by between-exemplar variability, a finding that at least does not automatically follow by the instantiation principle, given that there are indeed significant differences between the typicalities of the instantiations in each of the 20 subordinates.

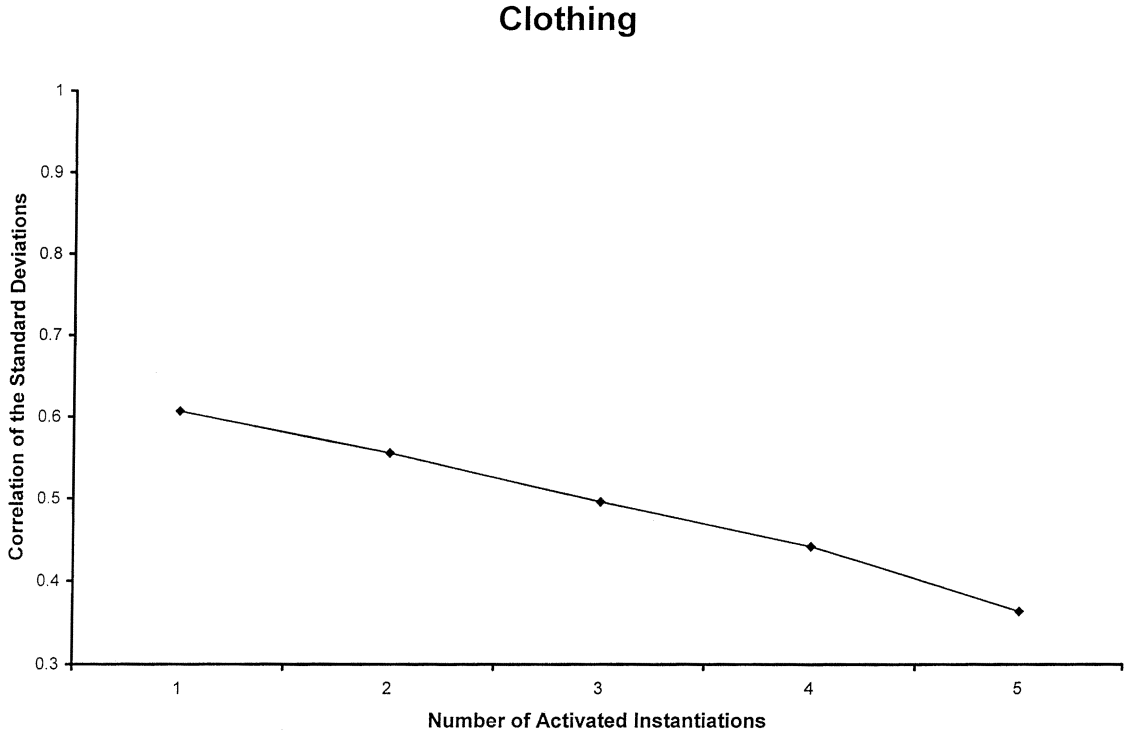


Figure 4. Correlations of observed and predicted standard deviations as a function of the number of activated instantiations in Experiment 2.

EXPERIMENT 3

Heit and Barsalou investigated two different simple categories (*food* and *animals*), besides two complex categories (*dangerous animals* and *small animals*). In terms of the concept hierarchy described by Rosch, Mervis, Gray, Johnson, and Boyes-Braem (1976), these concepts are situated at the super-superordinate level (see also Coley, Medin, & Altran, 1997; Tanaka & Taylor, 1991). Their ''subordinate categories'' (e.g., *fruit*, *vegetables*, etc. for *food*; and *birds*, *mammals*, etc. for *animals*) are situated at the superordinate level in Rosch et al.'s taxonomy. One may ask at what taxonomic level instantiation occurs. In the General Discussion section of their paper, Heit and Barsalou (1996) mentioned this subject, stating that it is not clear ''how far down the instantiation process goes'' (p. 444). For example, if subjects instantiate *fruit* with one or more instances (e.g., *apple*, *grape*, *orange*), do they then instantiate *apple*, *grape*, etc. with specific exemplars and/or kinds (e.g., *Granny Smith*, *Jonathan* for *apple*, and *Cabernet Sauvignon*, *Chardonnay* for *grape*)? Practically speaking, one may ask whether the instantiation principle can also be applied to predict typicalities within concepts defined at a hierarchically lower level. In this section, an experiment will be described that was similar to the experiments of Heit and Barsalou, but with the difference that typicalities were rated (and predicted) in categories defined at Rosch et al.'s superordinate level, where the subordinate categories are situated at Rosch et al.'s basic level. Instances of the subordinates are situated at a level below the basic level. As it cannot be evaluated on the basis of behavioural data whether people use exemplar or abstracted representations (Barsalou, 1990), we do not make any assumption about the nature of the instances.

Method

Participants. A total of 17 participants (undergraduate and graduate students of psychology) volunteered in the exemplar-production task; 20 first-year psychology students rated typicalities for the category *vehicles*; and 20 high-school students rated typicalities for the category *fruits*.

Materials. Two categories were selected that had to meet three criteria: (1) the cate-

gory had to be defined at the superordinate level of abstraction in Rosch et al.'s (1976) taxonomy, (2) the category needed to have at least seven subordinate categories defined at the basic level of abstraction in the same taxonomy (seven being the minimum number of subordinates that Heit & Barsalou, 1996, used in their experiments), and (3) these basic-level subordinates needed to have at least several lexicalised and well-known instantiations. Based on these three criteria, the categories *vehicles* and *fruit* were selected. For the exemplar-production task, the (subordinate) category cues were *bicycle*, *boat*, *car*, *cart*, *plane*, *spacecraft*, and *truck* for *vehicle*; and *apple*, *berry*, *cherry*, *grape*, *melon*, *nut*, and *pear* for *fruit*. For the typicality rating task, the rated categories were the 7 subordinates plus 180 unique instantiations that the exemplar-production group produced for *vehicles* and the 7 subordinates plus 83 unique instantiations that the exemplar production group produced for *fruit*.

Procedure. The procedure was like Experiment 1, except that the participants in the rating task judged the categories on a scale from 1 to 10 with respect to their typicality in *vehicles* or in *fruit*.

Results and discussion

As in the previous two experiments, two sorts of simulation processes were run: a simulation based on the assumption of only one single instantiation per subordinate, and a simulation based on the assumption of multiple instantiations. The simulation procedures were exactly the same as described in Experiment 1.

Prediction of the means. For *vehicles*, the correspondence between the observed means and the predicted means based on the single-instantiation simulations was very high: $r = .95$ ($p < .01$). (Given the estimated reliability coefficients of the observed and predicted measures, the maximum correlation, obtained by applying the correction for attenuation formula, was. 955.) The high correlation was not attributable to extreme values, as is illustrated by the perfect Spearman rank correlation for the same data: ($r = 1.00$, $p < .01$). For *fruits*, the correlation between the observed means and the predictions for the single-instantiation model was somewhat lower: $r = .70$ ($p < .05$), corresponding to a Spearman rank correlation of $r = .68$ ($p < .05$).

Correlations were again calculated between the means of the empirically obtained data and of the predictions from the multiple-instantiation model, based on the first n instantiations of the subordinates, with n varying from 2 to 5. Figure 5 shows the resulting correlations for both concepts as a function of n. For *vehicles*, the correlation decreased slightly as n increased. For *fruits*, the correlation increased from single instantiation ($n = 1$) to multiple instantiation with two activated instantiations, but decreased when more than two activated exemplars are assumed.

As in the previous two experiments, the alternative explanation (of a representation of the instantiations at the subordinate level) was evaluated by performing analyses of variance on the simulated data for each of the seven subordinate categories for *vehicles* and *fruits* using the single instantiation frequencies. This alternative interpretation, which we did not intuitively find very convincing for concepts like *food* and *clothing*, seemed somewhat more plausible for the lower-level concepts studied in Experiment 3. It might be that participants evaluate the typicality of different sorts of apples within the concept *fruits* just by activating a general representation of *apple*, thus not taking into account the differences (mostly in colour and in taste) between these different sorts of apples. The results yielded significant differences ($p < .01$) for all subordinates of both *vehicles* and *fruits*, except for the instantiations of *pear*, for which the typicalities did not differ significantly. Thus, as in Experiments 1 and 2, there were again substantial differences between the activated instantiations within all subordinates, arguing against an interpretation that assumes that the subordinate-level concepts determine the typicalities of their instantiations.

Prediction of the standard deviations. The correlation between the observed standard deviations and the single-instantiation predictions for *vehicles* was $r = .73$ ($p < .05$). The Spearman rank correlation was not significant ($r = .39$). For *fruits*, the corresponding correlation was only .35 (not significant).

Figure 6 shows the correlations of the standard deviations for both concepts as a function of n. Unlike in Experiments 1 and 2, the correlation for *vehicles* increased as n increased from 1 to 4, but then decreases as n gets larger than 4. For *fruits*, the (non-significant) correlation for the single

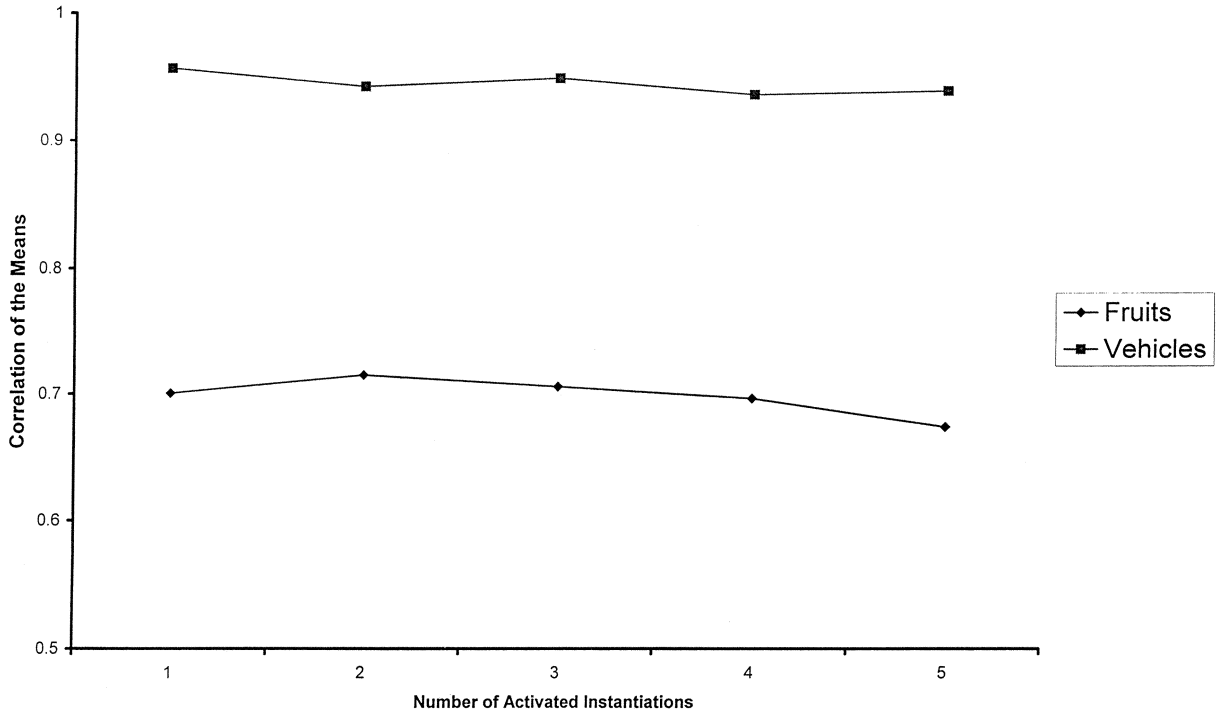


Figure 5. Correlations of observed and predicted typicalities as a function of the number of activated instantiations for the two concepts studied in Experiment 3.

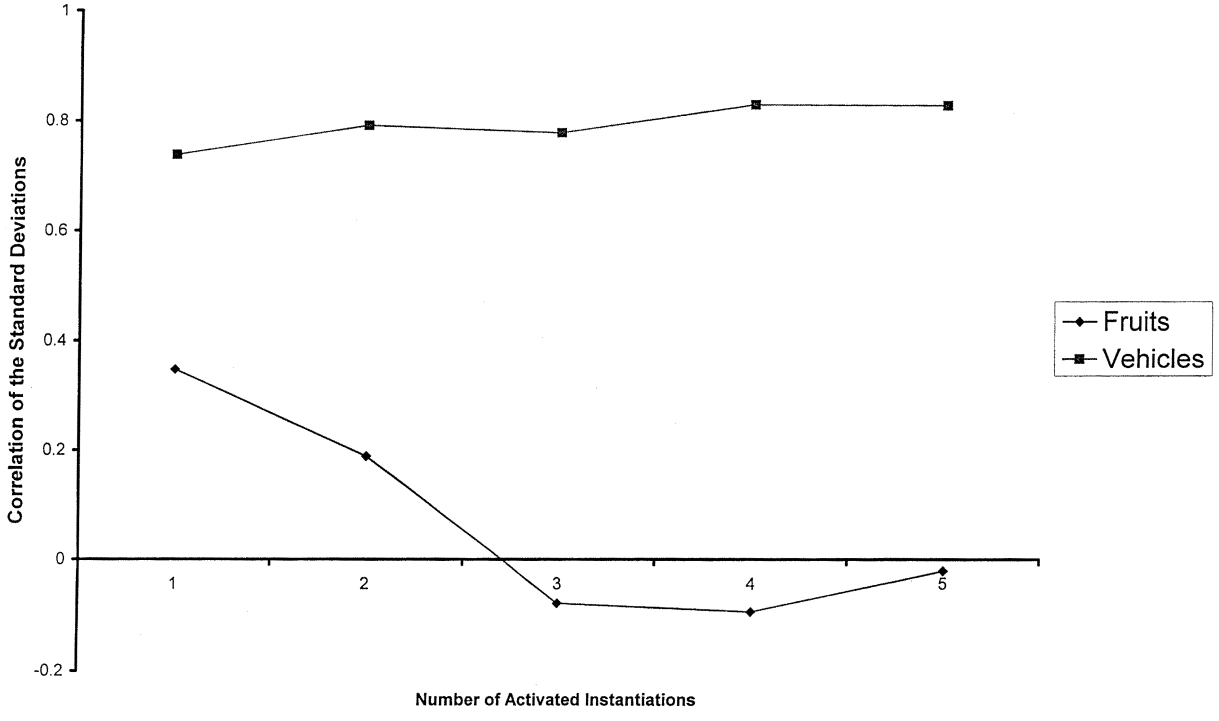


Figure 6. Correlations of observed and predicted standard deviations as a function of the number of activated instantiations for the two concepts studied in Experiment 3.

instantiation assumption still decreased if more instantiated exemplars are assumed. As in the previous two experiments, the standard deviations decreased in absolute size as the number of assumed instantiations increased from 1 to 5. As a consequence, the decreasing correlations for increasing numbers of instantiated exemplars for *fruits* might be an artefact due to restriction of range.

Analyses of variance were done for each of the seven subordinates of *vehicles* and of *fruits* to separate between-instantiation variability from within-instantiation variability. Both sources of variability were again correlated with the variances of the empirically obtained data, resulting all in non-significant correlations (.43 and −.36 for between-instantiation variability of *vehicles* and *fruit*, respectively, and .66 and .29 for within-instantiation variability of *vehicles* and *fruit*, respectively).

In conclusion, the results show that the instantiation principle predicted the means of the subordinates' typicalities well, even though these subordinates are defined at the basic level (as defined by Rosch et al., 1976), and thus the instantiations, which guide the prediction, are defined at a taxonomic level that is *lower* than the basic level. For *vehicles*, multiple instantiation did not result in better predictions than single instantiation, but note that the correlation was already at its maximum for single instantiation given the unreliability of the ratings. For *fruit*, a prediction based on two activated instantiations was somewhat better than the single-instantiation prediction. In general, it is not surprising that the number of activated instantiations decreased as the level of abstraction of the concepts studied decreased: Knowledge of such specific instances is most likely less elaborate than knowledge of more abstract concepts. (Even though there are many varieties of apples, most people's knowledge of different varieties is most likely rather limited.) The results for the standard deviations were, as in the previous experiments, less clear. We will elaborate on the predictions of the standard deviations next, in the General Discussion section.

GENERAL DISCUSSION

The results of the three experiments reported in this paper support the instantiation principle, proposed by Heit and Barsalou (1996). Experiment 1 replicated the findings of Heit and Barsalou in that high correlations were obtained for the means and the standard deviations of the observed and simulated distributions of typicality ratings for the subordinate categories of the concept *food*. Similarly high correlations were obtained in Experiment 2 for the concept *clothing*, where the correlations were based on a larger set of subordinate categories, which made the correlations more reliable.

In the first two experiments, predictions based on a single-instantiation assumption as described by Heit and Barsalou (1996) were compared with predictions based on multiple instantiations, with the number of activated instantiations varying from two to five. Correlations were used as an index of fit. Looking at the predictions for the means of the typicalities of the subordinate categories, multiple instantiations with two or three activated exemplars yielded somewhat better predictions, but the improvement was generally small. However, simulations based on the single instantiation principle almost invariably yielded somewhat better predictions for the standard deviations as compared to multiple instantiations. However, evidence was found that the lower correlation for multiple instantions might be the result of restriction of range. This plausible bias suggests one should not attach too much importance to the decreasing correlations for the multiple-instantiation predictions.

Experiment 3 tested whether the instantiation principle yielded equally good predictions for concepts lower in the semantic taxonomy (Rosch et al., 1976). The results indicated that both for *vehicle* and for *fruit*, mean typicalities and the corresponding standard deviations of the subordinate categories (which are defined at the basic level) can very well be predicted using simulated data that were generated according to the instantiation principle. When comparing single- and multiple-instantiation predictions, similar conflicting evidence as in the first two experiments was obtained. In the prediction of the mean typicalities of *fruit*, the correlation favoured multiple instantiation with two activated exemplars. *Vehicles* was the only investigated concept for which the correlation favoured single instantiation. In the prediction of the standard deviations, again the conclusion was not clear: sometimes single instantiation yielded better results, sometimes multiple instantiations yielded better predictions, but again these correlations might be biased by restriction of range. As in the first two experiments, the standard deviations decreased in

absolute size with an increased number of activated exemplars. However, it would not come as a surprise if instantiation were less elaborate when dealing with less abstract concepts. Despite the difficulties in judging whether single or multiple instantiation predicted the data better, both models yielded very good fits to the data.

The three experiments allow us to conclude that the instantiation principle, especially when eleborated to include more than one activated instantiation, yielded very good predictions of mean rated typicalities in concepts defined at the superordinate and the basic level. Interpreting the correlations of the standard deviations was more problematic for two reasons. First, restriction of range most likely biased the results of the comparison of the single-instantiation and the multiple-instantiation models. Second, after splitting up the variance of the simulated data into between-instantiation variance and within-instantiation variance, the high correlation obtained for the single-instantiation model was shown to be based on within-instantiation variance only, that is, on a correspondence of the empirically obtained variance with the variability due to inter-individual differences among subjects in rating particular instantiations. The simulated data, however, showed that the model usually also predicts significant correlations between the empirically obtained variances and the between-exemplar mean squares. There is no straightforward explanation based on the instantiation principle for the latter correlation not being found empirically.

A plausible explanation for the high correlation of the variance in the rated typicalities with the within-instantiation variability and the low correlation with the between-instantiation variability might be that this correlation corresponds mainly to inter-subject instability in the rated typicality of the exemplars. In other words, the within-instantiation variability reflects basically inter-individual differences in rating the typicality of the instances. If this is the case, then one might expect that the intersubject instability is larger for instances with lower typicality (with which the subjects are less familiar), and thus that the inter-subject instability (and therefore also the variance) is larger for concepts with more atypical instantiations. To investigate this hypothesis, the mean typicalities of the subordinate categories were correlated with the standard deviations of the corresponding typicalities. For the concept *food*, the means and the standard deviations of the subordinate's typicalities correlated $-.78$. Similar correlations for *clothing*, *vehicles*, and *fruit*, were $-.47$, $-.53$, and -93. All these values are significant at the .05 level, except for the correlation for *vehicles* ($p < .10$), which was only based on seven data points. This finding is compatible with the instantiation principle, but it is also compatible with a lot of other models of concept representation. Prototype models, for instance, which assume that the representation of a concept consists of directly stored features of the concept, also hold that less familiar concepts (of which subjects don't exactly know all features) are judged less reliable than typical concepts. The alternative hypothesis that was falsified in Experiments 1 to 3 is also compatible with this finding. Most important, however, is that the instantiation principle has no explanation for the fact that the variance in the empirical data does *not* correlate with the between-instantiation variability. We are inclined to conclude that the instantiation principle does a terrific job in predicting the mean typicalities of the subordinate categories, but that the simulated data do not reflect the second moment of the distributions due to several possible biases, some having to do with restriction of range, others with inter-rater instability due to unfamilarity.

The predictions for the means in the experiments reported in this paper provide some evidence that multiple instantiation yielded better predictions than single instantiation, although the difference was small. Two remarks should be added to this conclusion. First, the multiple instantiation hypothesis was tested using different numbers of activated instantiations (varying from two to five). The number of activated instantiations was always constant over the set of subordinate categories, but there is no reason to assume that this is necessary. When judging the typicality of *beverage* for the concept *food*, for instance, it is very well possible that only *Coca-Cola* is activated, whereas when rating the typicality of *fruit*, several instantiations are activated (or vice versa). For instance, for the concept *clothing*, we selected, out of the five predictions corresponding to one to five activated instantiations, the predicted typicality value for every subordinate category that best approaches the empirical value (in absolute value). These values correlate .95 with the empirically obtained means (which is a substantial improvement of the .93 found in Experiment 2, given that a correlation of .965 can be considered the maximum correlation taking into account the reliability estimates of the ratings). Second, the probabilities of the random

selection of the instantiations were based on the frequencies of the exemplars as first, second, third, etc. response in an instantiation-generation task. The only restriction was that the sampling was done without replacement. However, a more precise model would base the sampling on *conditional* probabilities, taking dependencies in the generation responses into account. For instance, given that a subject instantiates *apple* for the category *fruit* (as first instantiation, or as a subsequent instantiation), the response *pear* might have a high probability of being generated next, because of the association between apples and pears. We expect a model that incorporates conditional probabilities to yield better predictions. However, reliable estimates of such conditional probabilities require huge numbers of participants in the instantiation-generation task.

The results of Experiment 3 suggest that typicalities of basic-level concepts within superordinate-level concepts (in terms of the taxonomy presented by Rosch et al., 1976) are based on instantiations of the basic-level concepts. This finding raises the question whether typicalities at higher taxonomic levels (like the typicality of *fruit*, etc. in the concept *food*) are based on instantiations of these subordinates (like *apple*, *pear*, etc.), or on instantiations of these instantiations (like *jonagold apple*, *conference pear*, etc.). The direct results of the instantiation-generation task suggest the former, as participants who generate instantiations of *fruit* come up with answers like *apple* and *pear*, not with answers at a more specific level, like *jonagold apple* or *conference pear*, even though the instructions did not mention an intended level of abstraction for the responses. However, generating the verbal label ''apple'' as an instantiation of the concept *fruit* does not mean that information at the level of abstraction of *apple* is activated instead of a more concrete representation at the level of *jonagold apple* or even at the level of a specifically encountered *jonagold apple* purchased and eaten at a particular moment. The fact that typicalities of instantiations at the more specific level (of *jonagold apple*, for instance) still differ substantially suggests that these ratings are based on separate semantic representations. Just how far down does the instantiation principle go? As Heit and Barsalou (1996) argue, the procedure used in their paper and in ours does not allow us to answer this question. Further studies, using different paradigms, are needed for this question to be solved.

Finally, one might wonder how well the predictions of the instantiation model are as compared to prototype predictions. Storms et al. (2000) showed that rated typicalities can be predicted significantly better by the instantiation model than by prototype predictors calculated as in Hampton (1979) and in Rosch and Mervis (1975). The typicality rating task, however, differed from the typicality tasks described in Heit and Barsalou (1996) and in this paper. Although we did not gather the necessary data to construct similar prototype predictions for the tasks described in these experiments, the very high predictive correlations for the means presented in this paper, together with the results of Storms et al., suggest that prototype predictors would not outdo the instantiation predictors.

Manuscript received 16 July 1999
Manuscript accepted 18 November 1999

REFERENCES

Barsalou, L.W. (1990). On the indistinguishability of exemplar memory and abstraction in category representation. In T.K. Srull & R.S. Wyer (Eds.), *Advances in social cognition* (Vol. 3, pp. 61–88). Hillsdale, NJ: Lawrence Erlbaum Associates Inc.

Coley, J.D., Medin, D.L., & Altran, S. (1997). Does rank have its privilege?: Inductive inferences with folk biological taxonomies. *Cognition*, *23*, 73–112.

Hampton, J.A. (1979). Polymorphous concepts in semantic memory. *Journal of Verbal Learning and Verbal Behavior*, *18*, 441–461.

Hampton, J.A. (1993). Prototype models of concept representations. In I. Van Mechelen, J.A. Hampton, R.S. Michalski, & P. Theuns (Eds.), *Categories and concepts: Theoretical views and inductive data analysis*, (pp. 67–95). London: Academic Press.

Heit, E., & Barsalou, L.W. (1996). The instantiation principle in natural language categories. *Memory*, *4*, 413–451.

Jones, G.V. (1982) Stacks not fuzzy sets: An ordinal basis for prototype theory of concepts. *Cognition*, *12*, 281–290.

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

Lord, F.M., & Novick, M.R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

Medin, D.M., & Schaffer, M.M. (1978). Context theory of classification learning. *Psychological Review*, *85*, 207–238.

Medin, D.L., & Smith, E.E. (1984). Concepts and concept formation. *Annual Review of Psychology*, *35*, 113–138.

Nosofsky, R.M. (1986). Attention, similarity, and the identification–categorization relationship. *Journal of Experimental Psychology: General*, *115*, 39–57.

Nosofsky, R.M. (1988). Exemplar-based accounts of relations between classification, recognition, and typicality. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *14*, 700–708.

Rosch, E., & Mervis, C.B. (1975). Family resemblances: Studies in the internal structure of categories. *Cognitive Psychology*, *7*, 573–605.

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

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

Storms, G., De Boeck, P., Hampton, J.A., & Van Mechelen, I. (1999). Predicting conjunction typicalities by component typicalities. *Psychonomic Bulletin & Review*, *6*, 677–684.

Storms, G., De Boeck, P., & Ruts, W. (2000). Prototype and exemplar based information in natural language categories. *Journal of Memory and Language*, *42*, 51–73.

Tanaka, J.W., & Taylor, M. (1991). Object categories and expertise: Is the basic level in the eye of the beholder? *Cognitive Psychology*, *23*, 457–482.

Wisniewski, E.J. (1997). When concepts combine. *Psychonomic Bulletin & Review*, *4*, 167–183.

Copyright © 2003 EBSCO Publishing