Supporting materials

Pretest

This pretest was conducted to test whether sets whose members are completely identical or completely distinct on each salient feature are rated as better than sets for which this is not the case (i.e., some members are the same, but some differ, on a salient feature).

Method

Participants (*N* = 81, 72 females, M_{age}= 19.28) were students from Tilburg University who participated in return for course credit. Participants were shown nine pictures depicting sets of five cartoon dinosaurs. They were instructed that “some sets of things feel better than other sets of things. Please indicate how good the set feels and (if applicable) how the set can be improved”. The instructions were intentionally left vague because we wanted to test whether people have a notion of what a “good” set is without defining what good was and thereby creating demand effects. For each set we created a “good” version for which all dinosaurs were the same—or all differed—on the attributes of color and type (i.e., shape), and a “bad” version for which this was not the case (i.e., some dinosaurs were of the same color/type, while others differed; see Figure S1).

Participants were randomly shown either the good or the bad version of each of the nine sets. For each set, they indicated how good it felt on a 7-point scale (1 = not at all good, 7 = very good). They were also asked if, and if so how, the set could be improved.

Results

We excluded four participants from analysis. One participant failed to mark down the correct trials, the other three gave responses to the questions on how to improve the set that were classified as “weird” for all 9 trials (for example, one participant suggested for every set giving the dinosaurs camouflage colors, the other suggested making them look more angry so they appeared more dangerous). Including these participants from the analysis did not lead to any meaningful differences in the ratings of the goodness of the sets (all test statistics are still

$p < .001$).

Rated goodness of set

Participants indeed judged the sets following the good sets as better than the bad sets: Wilk's $\Lambda = .23$, $F(9,67) = 25.40$, $p < .001$, $\eta^2 = .77$. For univariate results, see Table S1.

Suggested improvements

When coding the responses to the open-ended question on how to improve the set, we ran into an unexpected regularity. Although we intentionally varied the dinosaurs on two aspects (color and type) we unintentionally also varied them on two other aspects. One of the dinosaurs was a pterodactyl (meaning it could fly), and one of the dinosaurs was facing the right where others were all facing the left. Because changing both these aspects was suggested as an improvement several times, we coded the responses in the following categories:1) no improvement suggested; 2) improvements following the our proposed rules (in the example bad set this would be changing the color of one of the purple dinosaurs to a color not used yet or removing one of them); 3)other improvements following rules (in good set example flipping the red dinosaur over so all are facing the same direction); 4) expanding the set; and 5) other (e.g., spacing the dinosaurs more evenly).

For bad sets, almost 90% of the participants suggested that the set could be improved in a way that follows the proposed rules for good sets (see Figure S2). This is further evidence that there indeed seem to be general principles that determine whether a set feels good or not. These rules were also apparent for good sets. Here a lot of people suggested (more than 20%) that the set should be expanded by adding another dinosaur according to the hypothesized principles (for example for FigureS1; adding an orange dinosaur of a different type). Since the dinosaurs unintentionally differed on other aspects as well, we also looked at differences between the conditions for suggestions relating to these differences. Interestingly, for bad sets only 2.9% gave suggestions pertaining to these differences (e.g., Figure S1, removing the yellow flying dinosaur since it is the only one that can fly) whereas for good sets 14.7% gave such suggestions, suggesting that when differences in type or color were present, these were more salient than the other aspects (left/right facing and flying/non-flying).

Experiment reported in General Discussion

Method

Participants (N = 204, 50females, M_{age}= 27.41) recruited on Amazon.com's Mechanical Turk, were randomly assigned to one of four conditions in which they made a choice between two sets of mugs (see Figure 5 in article). In each condition, the only difference between the two choice options (the sets between which the participants chose) was a green vs. an orange mug. In the two *all-similar* conditions the green and orange mugs were accompanied by 3 more green mugs in the *all-similar-green* condition, and by 3 more orange mugs in the *all-different-orange* condition. In the two *all-different* conditions, we added a blue, pink and green mug to both options in one condition, and a blue, pink and orange one in the other condition. Based on set-fit, we expected that adding three green would increase the green mug's choice share. Adding three orange mugs should increase choice for the orange mug. In the *all-different* conditions set-fit predicts that adding the single green mug (plus the pink and blue one) decreases the green mug's choice share, and adding the single orange one (plus the pink and blue) decreases the orange mug's choice share.

Results

In the all-similar orange condition, 62.3% (33/53) of the participants chose the set including the orange mug. In the all-similar green condition, only 39.6% (19/48) chose the set including the orange mug; $\chi^2(1, n= 101) = 5.19$, $p = .03$, $\varphi = .23$.

In the all-different condition in which a pink, blue and green mug was added, 94.1% (48/51) chose the set including the orange mug. When a pink, blue and orange mug were added, only 15.7% (8/51) chose the set including the orange mug; $\chi^2(1, n= 102) = 63.4$, $p< .001$, $\varphi = .79$.

Figure S1. “Good” and “bad” versions of Set 1. On the left is the bad set: All dinosaurs are of different types, but some share the same color whereas others do not. On the right is the good set: All dinosaurs completely differ both on type and color.

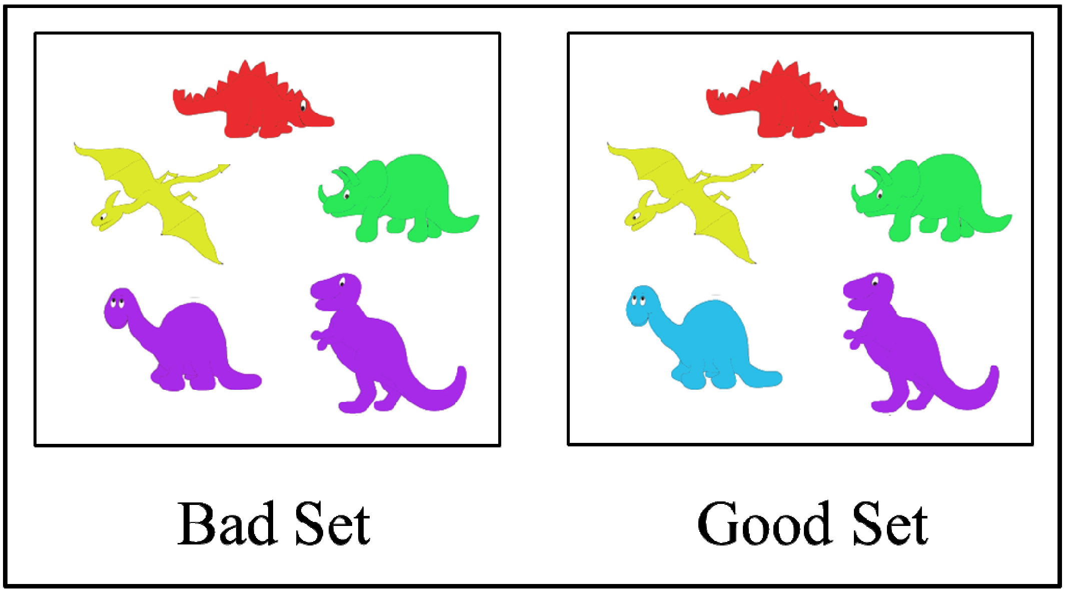

Figure S2. Improvements (%) suggested for good and bad sets.

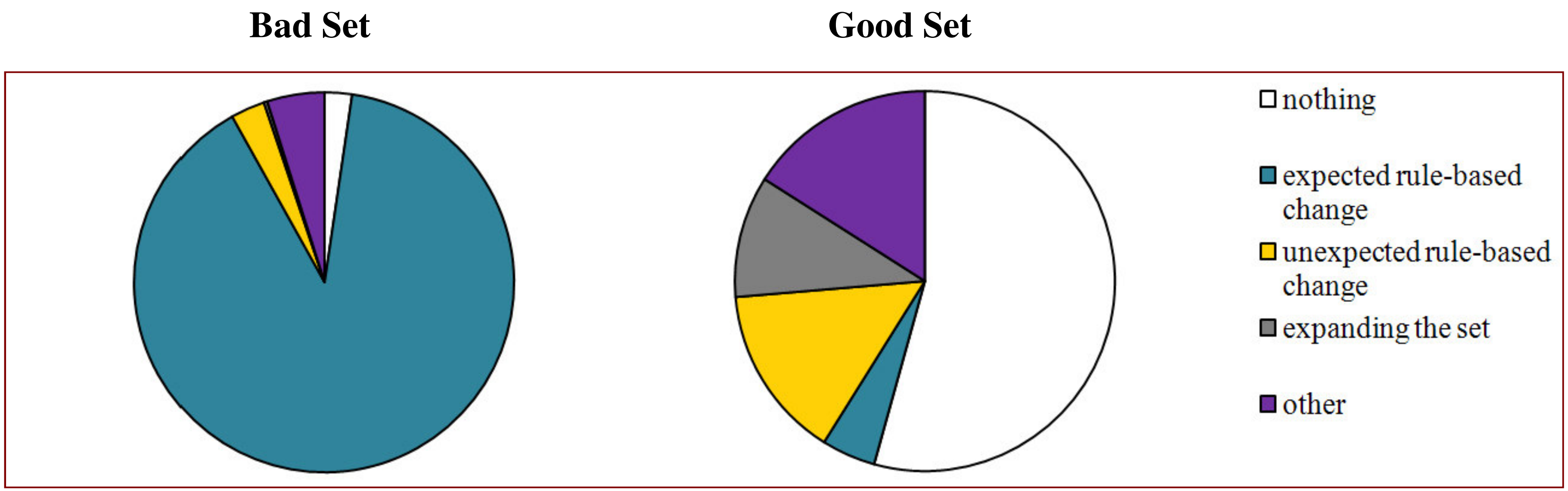

Table S1. Means and standard deviations of rated goodness for the good and bad versions of each of the 9 sets.

	Good set		Bad set			
	M	*(SD)*	*M*	*(SD)*	*p*	η^2
Set 1	5.26	(1.61)	3.31	(1.34)	< .001	0.31
Set 2	5.74	(1.74)	4.39	(1.70)	< .001	0.14
Set 3	5.77	(1.37)	4.42	(1.31)	< .001	0.21
Set 4	5.97	(1.50)	3.62	(1.35)	< .001	0.41
Set 5	6.38	(1.14)	3.82	(1.54)	< .001	0.48
Set 6	6.58	(0.72)	3.67	(1.53)	< .001	0.60
Set 7	6.26	(0.94)	4.08	(1.57)	< .001	0.42
Set 8	5.59	(1.62)	3.95	(1.27)	< .001	0.25
Set 9	6.58	(0.76)	3.38	(1.51)	< .001	0.64