Product Lineups: The More You Search the Less You Find

Web Appendix

This web appendix contains the stimuli and instructions used in the experiments. It also includes supporting analyses for certain studies and a single-paper meta-analysis.

WEB APPENDIX A

Preliminary survey instructions and results

We recruited participants through Amazon Mturk (N = 448; mean age = 34.94; 44.4% women) as well through social media (N = 73; mean age = 27.1; 31.5% women), to increase the generalizability of the findings. Participants saw a single question and chose their response among six alternatives:

“Have you ever been in a situation where you saw a product somewhere, but were not sure of the exact brand or model, and later tried to find the product in online or offline stores, trying to identify it visually among other products that looked similar?”

a. Yes, it happens to me frequently
b. Yes, it happens to me occasionally
c. Yes, it has happened to me, but not frequently
d. It hasn’t happened to me personally, but I can image it happening to other people, or even to me in the future
e. It hasn’t happened to me and I don’t see it ever happening
f. I do not understand the type of situation you described

Note: Mturk participants were also asked to briefly describe their experience.

The results did not differ across samples ($\chi^2(8) = 8.497$, $p = .386$).

	MTurk (N = 307)	Non-MTurk (N = 73)	MTurk replicate (N = 141)	Total (N = 491)
Yes, it happens to me frequently	11.7%	8.2%	13.5%	12.0%
Yes, it happens to me occasionally	39.1%	41.5%	41.8%	40.3%
Yes, it has happened to me, but not frequently	40.7%	41.1%	29.1%	38.7%
Subtotal	**91.5%**	**91.8%**	**84.4%**	**91.0%**
It hasn't happened to me personally, but I can image it happening to other people, or even to me in the future	7.2%	6.8%	9.2%	6.9%
It hasn't happened to me and I don't see it ever happening	1.3%	1.4%	6.4%	2.0%
I do not understand the type of situation you described	0%	0%	0%	0%

Sample responses:

"*An acquaintance had a charging brick that worked very well for a fast charge. I didn't get the brand or model number but I remembered what it looked like and found it on Amazon.*" (Female, 50)

"*Sometimes I see a product through a post on Facebook, but the brand isn't mentioned. When I look it up, I find a lot of similar products, usually they're knockoffs so they look a lot alike but aren't made as well...*" (Female, 25)

"*I wanted a coffee maker and I remember seeing one at a store that looked great but I didn't have the money for it at the time. I tried looking it up on Amazon and what I found was similar but not quite what I saw at the store. I ended up getting the similar product on Amazon, not what I seen at the store, but something similar and cheaper.*" (Male, 22)

"*I saw a handbag that I liked when I was on the subway, so I searched for it online. I used the style, color, and pattern to try to find what I was looking for and I was able to find it.*" (Female, 29)

"*I was looking for a pair of Adidas shoes and I didn't know they were Adidas because they didn't have the logo on the side. It took me a while but I started finding shoes related to them and then found them.*" (Male, 22)

"*It was a handbag that I saw someone wearing. I googled the description of the bag trying to find out where to purchase it. I never found the product and assumed it was not currently being sold anymore.*" (Female, 31)

"*I saw a backpack in a movie that I thought was really cool. I couldn't see what the company name of it was in the movie so I had to spend an hour or two on Google trying to describe it and looking in the picture section.*" (Male, 29)

Note: Additional responses are available from the authors upon request.

WEB APPENDIX B

Participants’ identification accuracy for each sequence of lineup decisions

CORRECT IDENTIFICATION RATES IN STUDY 1

Sequence	1st	2nd	3rd	4th	5th	6th	7th
2-lures	76.9%	71.2%	65.4%*	X	X	X	X
6-lures	84.8%	65.2%	87.0%	91.3%	93.5%	84.8%	39.1%*

asterisk (*) denotes the correct target

WEB APPENDIX C

Supplementary materials for the follow-up study of study 1

1. Study procedure and results (see next section for an illustration)

One hundred fifty-eight participants (M_{age} = 32.8, 50.6% women; excluding 21 participants who did not follow the instructions, although including them in the analyses did not significantly alter the results) were recruited from MTurk. Participants were randomly assigned to one of two between-subjects conditions (2 vs. 6 lures).

First, participants saw a picture of a pair of headphones and were asked to imagine that they saw someone wearing it recently (see next section). They were further told that the pair of headphones shown in the picture was known to have high sound quality, and that they were thinking of buying the exact same headphones for themselves. On the next screen, the picture of the headphones disappeared, and participants imagined searching online for the exact product they had seen, and that they were not willing to settle for similar looking products. This was done so to prevent participants from deliberately memorizing the product image.

Next, all participants went through a filler task used in a previous research (i.e., reading a short essay on dolphins; Sela and Shiv 2009), which is also used in study 2.

Third, participants were told that they were going to see a number of headphone products as they would when they search for the product online. They were remined that their goal was to identify the exact product they were shown earlier.

To hold constant participants' perceptions of the likelihood of ever encountering the correct target, we told participants in both conditions that there were fifteen similar headphones overall, but only one of them was the correct product. We did this to give participants a concrete benchmark for the likelihood of encountering the true target on any given trial (i.e., 1/15).

As in study 1, those in the 6-lures (2-lures) condition saw six (two) lure headphones in a random order, each on a separate page, with a binary response question for each product asking whether it was the item for which they were looking. After the lures came the correct target they saw at the beginning of the survey. Participants indicated whether it was the correct product, which served as the dependent measure.

Results

The results replicated study 1. Participants who saw six lures before encountering the correct target were significantly more likely than those who saw two lures to make a false-negative judgment, erroneously rejecting the correct target (6 lures: 61.3% vs. 2 lures: 82.1%, $\chi^2(1) = 8.39$, $p = .004$, *Cohen's d* = .46). The effect was robust when only those participants who correctly rejected all the lures in the lineup were included in the analysis (6 lures: 65.6% vs. 2 lures: 85.1%, $\chi^2(1) = 6.62$, $p = .01$, *Cohen's d* = .47). These results suggest that our threshold escalation effect is unlikely to be driven by base-rate updating. Similar to study 1a, the lures were correctly rejected 86.6% of times

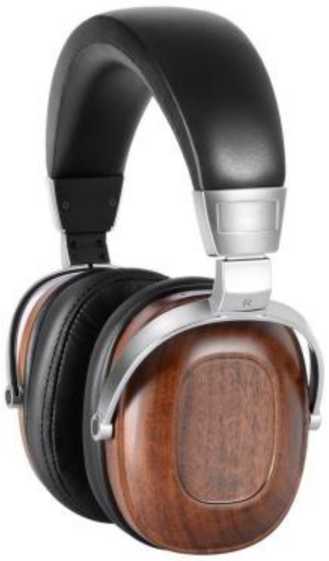
The mere-inaccuracy account was also not supported, given that the accuracy for the last lure was not different across the two conditions (6 lures: 98.4% vs. 2 lures: 98.5%, $\chi^2(1) = 0.004$, $p = .947$). While we did not replicate the increase in accuracy for those in 6-lures condition, we believe that the lack of difference reflects a ceiling effect, potentially driven by a use of different stimuli. However, the very high accuracy itself renders the inaccuracy account doubtful. The recognition rates for each order of judgments participants made is shown in the below table.

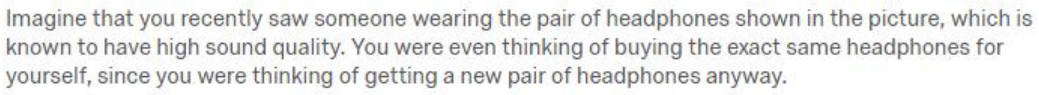
CORRECT IDENTIFICATION RATES IN STUDY 1 FOLLOW-UP

Sequence	1st	2nd	3rd	4th	5th	6th	7th
2-lures	87.2%	93.6%	82.1%*	X	X	X	X
6-lures	85.5%	92.5%	93.8%	96.3%	93.8%	96.3%	61.3%*

asterisk (*) denotes the correct target

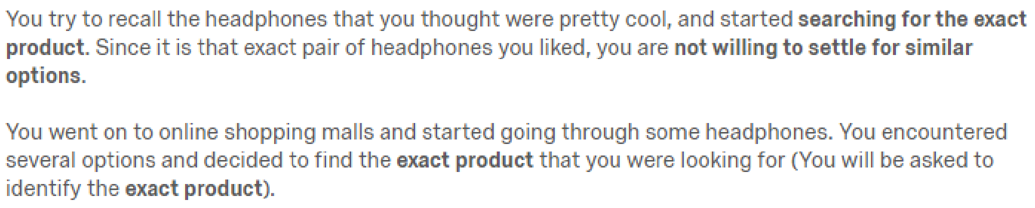
2. Illustration of the study procedure





Imagine that you recently saw someone wearing the pair of headphones shown in the picture, which is known to have high sound quality. You were even thinking of buying the exact same headphones for yourself, since you were thinking of getting a new pair of headphones anyway.

[Page break]



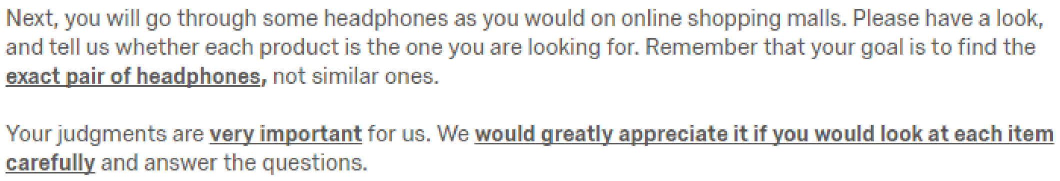
You try to recall the headphones that you thought were pretty cool, and started **searching for the exact product**. Since it is that exact pair of headphones you liked, you are **not willing to settle for similar options**.

You went on to online shopping malls and started going through some headphones. You encountered several options and decided to find the **exact product** that you were looking for (You will be asked to identify the **exact product**).

[Page break]

Filler task borrowed from Sela and Shiv (2009; reading a short article on dolphin culture).

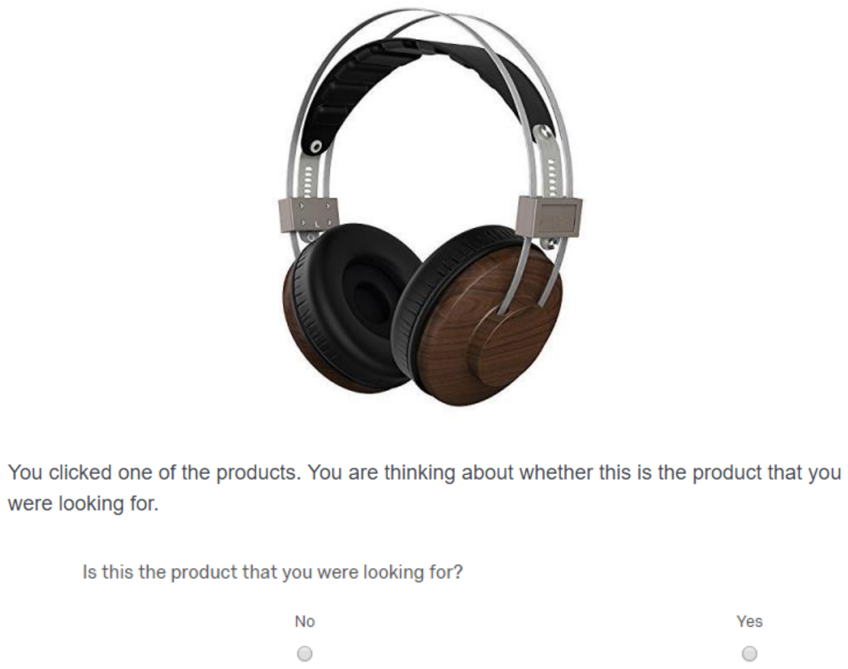
[Page break]



Next, you will go through some headphones as you would on online shopping malls. Please have a look, and tell us whether each product is the one you are looking for. Remember that your goal is to find the **exact pair of headphones,** not similar ones.

Your judgments are **very important** for us. We **would greatly appreciate it if you would look at each item carefully** and answer the questions.

[Page break]



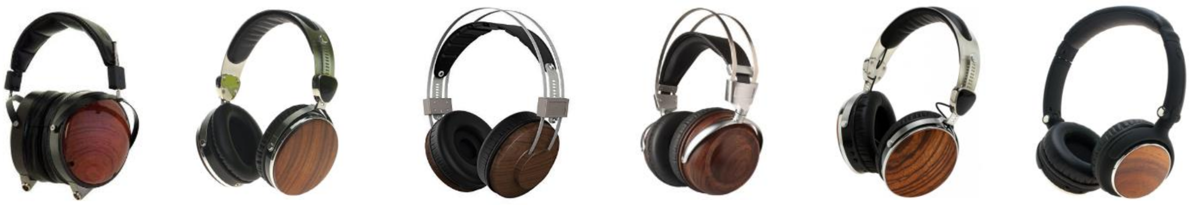
You clicked one of the products. You are thinking about whether this is the product that you were looking for.

Is this the product that you were looking for?

No ○ Yes ○

This procedure repeated five (one) additional times for those in 6-lures (2-lures) condition, randomly drawing one of the following sets of similar-but-not-identical headphones on each separate page. The correct target was presented after the lures seamlessly, and participants answered the same identification question.

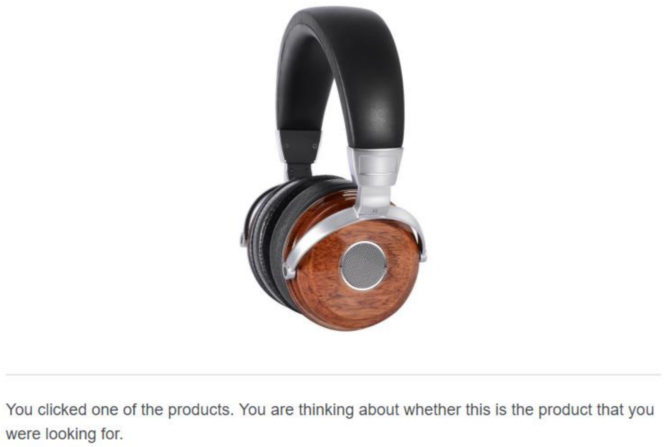
3. Set of 6 lures used in the study



WEB APPENDIX D

Supplementary materials for study 2

1. An example of simulated mismatch judgment



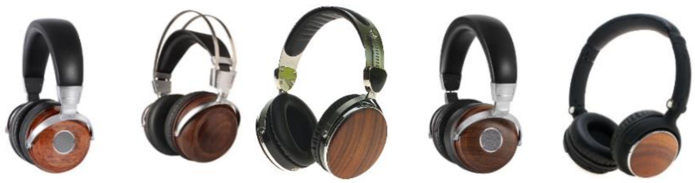
You clicked one of the products. You are thinking about whether this is the product that you were looking for.

After 5 seconds, the following words appear on the same screen:

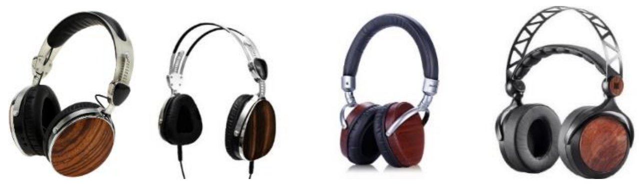
“Although it seems pretty similar, you realize that this is not the product you were looking for.”

2. Set of 9 lures used in the study

- Headphones presented to participants in 5-lures condition:

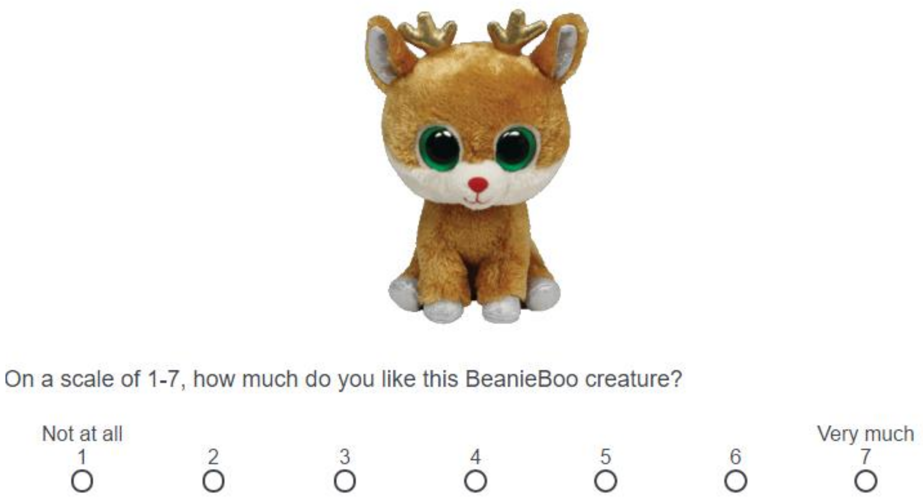


- Four additional headphones for 9-lures condition:



3. Filler items used in the 5-lures condition

In the 5-lures condition, participants first saw the following four filler items, and then saw five incorrect headphones randomly drawn among the nine shown above to match the number of judgments with 9-lures condition.



On a scale of 1-7, how much do you like this BeanieBoo creature?

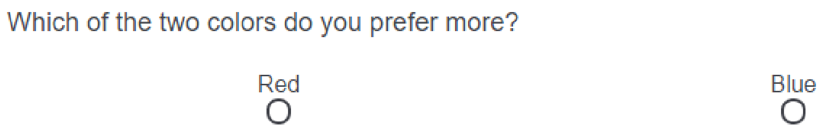
Not at all 1	2	3	4	5	6	Very much 7
○	○	○	○	○	○	○

[Page break]

How often do you dine out?

Not very often ○ ○ ○ ○ ○ ○ ○ ○ ○ Very Often

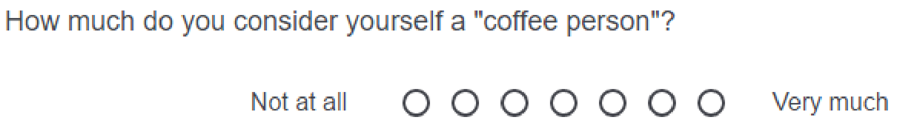
[Page break]



Which of the two colors do you prefer more?

Red	Blue
○	○

[Page break]



How much do you consider yourself a "coffee person"?

Not at all ○ ○ ○ ○ ○ ○ ○ Very much

WEB APPENDIX E

Follow-up studies procedures and results (study 2)

1. Testing an alternative explanation based on confidence

Two follow-up studies (combined N = 736) examine alternative accounts based on confidence. One may wonder if seeing more lures made participants feel less confident in identifying the correct target, which may have led to an increased likelihood of erroneously rejecting the correct target. On the other hand, one might expect those with a higher threshold (i.e., after seeing more lures) to be *more* confident in their decision than those with a lower threshold, given that an item is dismissed/accepted at higher standards. Despite the divergent predictions, we tested these accounts by directly measuring participants' confidence in their decision. We note, however, that according to our conceptualization, having a higher threshold means being more conservative and prudent, which does not necessarily imply increased or decreased confidence.

The procedure of the first follow-up study was similar to that of study 2, except that we varied the number of lures participants screened through (2 or 6 instead of 5 or 9) and the product category (sweatshirts instead of headphones) for generalizability. Three hundred and ninety-three undergraduate students (M_{age} = 20, 54.5%; none failed to follow instructions) participated in this study for a course credit. After participants screened through either 5 or 9 lures and made their identification decision for the correct target, we asked them to indicate how confident they felt about their decision (1 = *not at all confident*, 9 = *very confident*). Consistent with our prior findings, we expected false rejection of the true target to be higher in the 6-lures condition, but our conceptualization does not predict a difference in confidence across the two conditions.

The threshold escalation effect was replicated (6 lures: 73.5% vs. 2 lures: 83.2%; $\chi^2(1)$ = 5.54, p = .019, *Cohen's d* = .24), but we found no difference in confidence ($M_{\text{six lures}}$ = 7.07 vs. $M_{\text{two lures}}$ = 7.16; $F(1, 391)$ = 0.3, p = .586). Furthermore, our effect holds when controlling for confidence (p = .019), and confidence did not interact with the number of sequence on identification accuracy (p = .479).

The second follow-up study recruited 343 participants from MTurk (M_{age} = 37.7, 52.5%; excluding twenty-five participants who failed to follow instructions). The procedure was identical to that of study 2, except that we asked the same measure of confidence described above.

Again, although screening through a greater number of lures increased the likelihood of falsely rejecting the correct target when it appeared (9 lures: 59.8% vs. 5 lures: 77.5%; $\chi^2(1)$ = 15.51, p < .01, *Cohen's d* = .39), there was no difference in confidence between the two conditions ($M_{\text{nine lures}}$ = 6.68 vs. $M_{\text{five lures}}$ = 6.99; $F(1, 341)$ = 2.47, p = .12). Controlling for confidence did not significantly alter the effect (p = .001), and confidence did not interact with our manipulation (p = .981).

Taken together, the results from the two follow-up studies with participants from both MTurk and the lab suggest that confidence in judgment after seeing more lures is less relevant to the threshold escalation effect in our experiments.

2. Testing an alternative explanation based on doubt

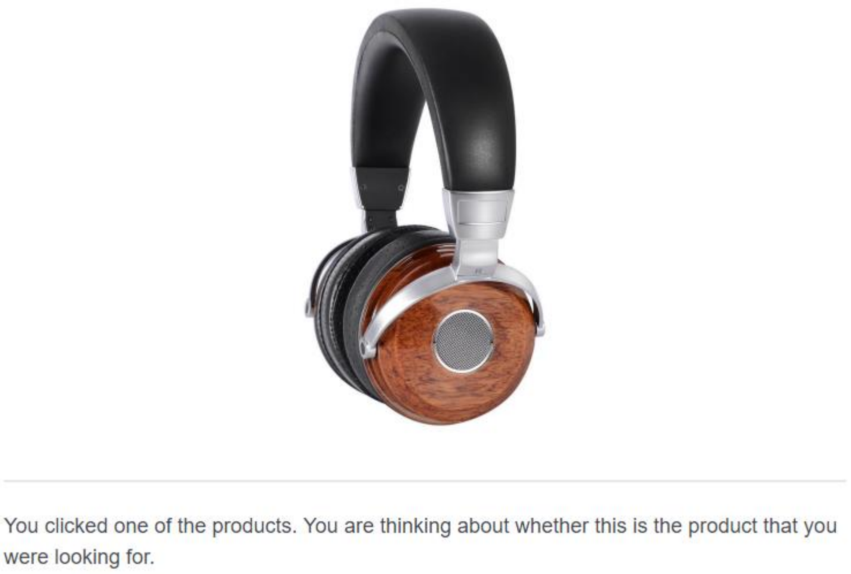
An alternative explanation may be that showing a greater number of lures may have somehow lead participants to doubt that the correct target will ever appear in the sequence, which in turn led them to doubt that the true target was in fact correct (independent of matching threshold). To address this concern, we directly measured perceptions of doubt in this follow-up study.

Four hundred and thirty-one participants from MTurk (M_{age} = 36.9, 55.5% women; excluding 16 participants who did not follow the instructions; including them in the analyses did not significantly alter the results) were randomly assigned to either a five or a nine lures condition. The procedure of this follow-up study was identical to that of study 2, except that participants responded to a doubt measure after the lures rather than making an identification decision for the correct target. Specifically, after screening through five or nine lures (depending on condition) and simulating making mismatch judgments for all of them, participants were asked “How likely are you to encounter the correct product soon?” (1 = *very unlikely,* 7 = *very likely*). The results showed that the magnitude of doubt did not differ between the two condition ($\text{M}_{\text{nine lures}}$ = 4.26 vs. $\text{M}_{\text{five lures}}$ = 4.38; *F*(1, 429) = 0.757, *p* = .385). Together, the results from this high-powered study suggests that participant doubt is an unlikely alternative explanation.

WEB APPENDIX F

Supplementary materials for study 3

1. Social comparison cue manipulation through simulated rejection paradigm



You clicked one of the products. You are thinking about whether this is the product that you were looking for.

After 5 seconds, the following words appear on the same screen for stringent (lax) condition, where the percentage of participants who presumably recognized the lures was varied:

> “Although it seems pretty similar, you realize that this is not the product you were looking for. 83% of people thought that this was (not) the right one in another survey.”

This procedure repeated five or nine times, depending on the condition, randomly drawing one of the following sets of similar-but-not-identical headphones on each separate page.

2. Study 3 pretest procedure and results

The objective of this pretest was to ensure that our lax social comparison cue indeed made participants feel that their threshold was already high enough compared to when the cue in the control condition was provided. To do so, we recruited seventy-six participants from MTurk, and randomly assigned them to either lax social comparison cue or control condition. The design of the pretest was identical to that of study 3, except that all participants only saw one randomly chosen lure, instead of 5 or 9 lures. Specifically, participants first imagined that they were looking for an exact pair of headphones, and then simulated rejecting a single, randomly chosen lure product. Those in the lax cue (control) condition were also told participants that 83% of participants in another survey incorrectly (correctly) thought that the lure was (not) the product they were looking for. Since the aim of this pretest was to gauge the effectiveness of this manipulation, a single feedback was deemed sufficient.

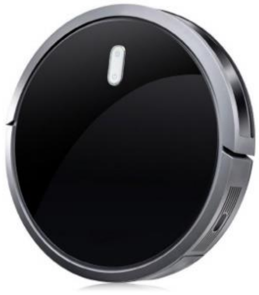
Afterwards, we asked 3 questions to measure participants' perceptions of how conservative their threshold is in different ways. The first questions asked whether other people's correct response rate was lower than expected, which is a precursor for participants to feel that their threshold is already relatively conservative ("Which of the following best describes what you think about other people's response?", 1 = *people are less accurate than I expected*, 7 = *people are more accurate than I expected*). The results showed that participants in the lax cue condition felt that people in the other survey were performing less accurate than expected ($M_{\text{lax cue}}$ = 2.25 vs. M_{control} = 4.93; $F(1, 74) = 61.35$, $p < .001$, *Cohen's d* = 1.80). The second measure directly asked whether participants felt that they were more discerning than others ("Compared to the participants in the other survey, I feel that I am more accurate", 1 = *strongly disagree*, 4 = *neither agree nor disagree*, 7 = *strongly agree*). As we intended, the lax cue made participants believe that they already have a discerning threshold ($M_{\text{lax cue}}$ = 5.67 vs. M_{control} = 4.98; $F(1, 74) = 4.56$, $p = .036$, *Cohen's d* = .49). Lastly, we asked how likely they would be to distinguish the lure products, which also captures the extent to which participants believe they are discerning ("Compared with our prior participants, how well do you think you can correctly spot incorrect product, like the one you saw on the previous page?", 1 = *I would be worse than the majority of* *participants in the previous survey*, 7 = *I would be better than the majority of participants in the previous survey*). The results were consistent with the objective of the manipulation, such that those in the lax cue condition perceived themselves to be more likely to be correct ($M_{\text{lax cue}}$ = 5.92 vs. M_{control} = 5.38; *F*(1, 74) = 4.67, *p* = .034, *Cohen's d* = .50). The converging evidence thus suggest that the lax cue manipulation leads participants to believe that they already have a relatively stringent threshold.

WEB APPENDIX G

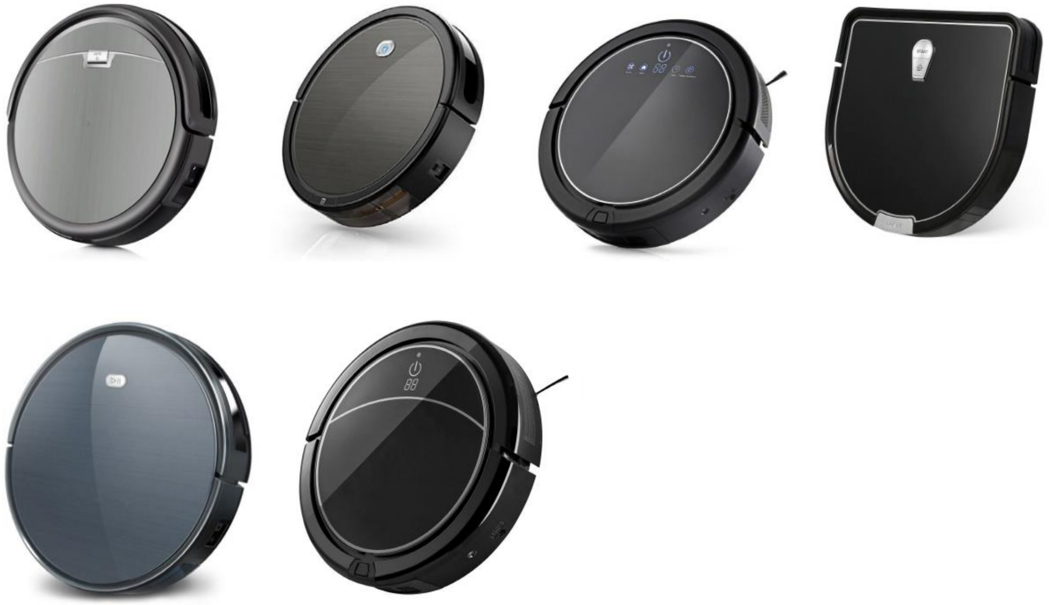
Supplementary materials for study 4

1. Robot vacuum cleaner stimuli used in study 4

- Correct Target:

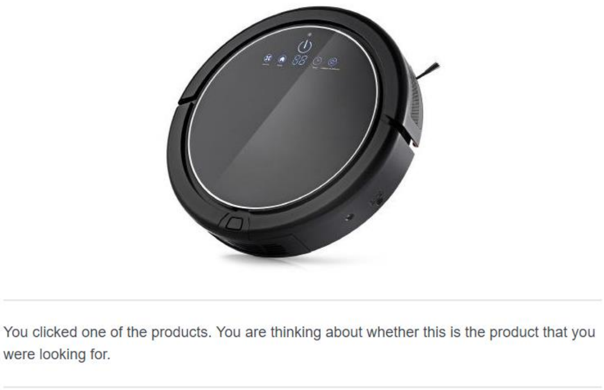


- Lures (2 or 6 shown randomly depending on the condition participants were assigned to):



2. Illustration of the expected difficulty manipulation used in study 4

- High expected difficulty condition

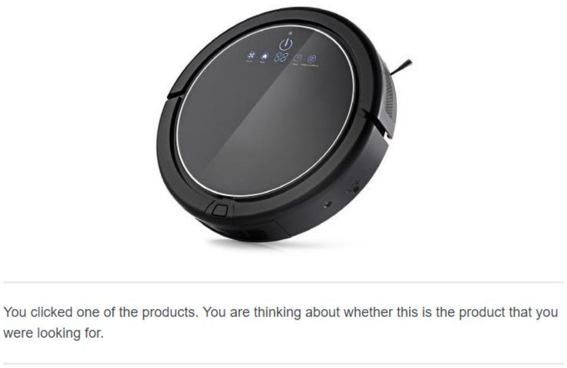


You clicked one of the products. You are thinking about whether this is the product that you were looking for.

After 5 seconds, the following words appear on the same screen:

“Although it seems pretty similar, you realize that this is not the product you were looking for. Google algorithm estimates that this product shares 83% of design features with the product you are looking for.”

- Baseline expectation condition



You clicked one of the products. You are thinking about whether this is the product that you were looking for.

After 5 seconds, the following words appear on the same screen:

“Although it seems pretty similar, you realize that this is not the product you were looking for. Google algorithm estimates that this product shares only 63% of design features with the product you are looking for.”

3. Study 4 pretest procedure and results

The high expected difficulty manipulation implemented in study 4 was designed to elevate participants' threshold by telling them the lures are objectively very similar to the true target. That is, to the extent that participants expect the recognition task to be difficult, they should adopt a stringent matching threshold (Benjamin and Bawa 2004). We examined whether the providing such additional information indeed led participants to believe the search process to be more difficult in this pretest.

The design of the pretest was identical to that of study 4, except that all participants screened through two randomly chosen lures, because we primarily predicted the high expected difficulty manipulation to increase false rejection rate of the correct target in the shorter sequence condition in study 4 (i.e., 2-lures condition). As in study 4, participants in the high expected difficulty condition saw two lure robot vacuum cleaners, and were told that a Google algorithm estimates the lures to share 89% and 91% of design features with the correct target, respectively. Those in the baseline condition saw the same lures, but were told that they were 61% and 69% similar to the correct target, respectively. Thus, the only difference between the two conditions was the estimated similarity by the algorithm.

After rejecting the lures, participants responded to three measures of expected difficulty of recognizing the correct target ("If we asked you to identify the product you saw earlier among these similar looking products, how likely do you think you might get confused?", 1 = *not at all likely*, 7 = *very likely*; "How easy do you think it will be to identify the correct one you saw initially, if you saw it?", 1 = *not at all easy*, 7 = *very easy*; "Given how similar the incorrect products are, how difficult would it be to tell the products you saw earlier among all the similar-looking products?", 1 = *not at all difficult*, 7 = *very difficult*), with the second measure reversecoded. We averaged the three measures to form an index of expected difficulty (*Cronbach's α* = .901).

In line with our prediction, participants in the high expected difficulty condition expected the recognition task to be significantly more difficult than those in the baseline condition ($M_{\text{high difficulty}}$ = 4.39 vs. M_{baseline} = 3.46; *F*(1, 77) = 7.89, *p* = .006, *Cohen's d* = .63), despite the fact that participants in both conditions saw the same lure products. The results thus provide support for the validity of the manipulation.

WEB APPENDIX H

Supplementary materials for study 5

1. Procedure and results for study 5

Method

Three hundred and seventy-one participants were recruited through MTurk (M_{age} = 36.5, 60.6% women; no participant failed to follow instructions). They were randomly assigned into a condition in a 2 (lures: 2 vs. 6) x 2 (distinctive cue: present vs. absent) between-subjects design. We asked participants to imagine that they were looking for a pair of headphones, as in study 2. However, participants made spontaneous correct/incorrect judgments for each option presented in the lineup, rather than simulating the process.

To bolster generalizability, we used a slightly modified procedure for the lineup. After reminding participants that their goal was to find the exact pair of headphones shown at the beginning of the study, we told them that they would see the search results generated by several online shopping platforms, based on a description of the target headphones they had seen. We also told participants that the correct product may or may not be among the options provided since search engines are often inaccurate.

On the next page, all participants saw eight hyperlinks, all labeled ‘click to view the product’ (see below section). We reminded participants that it was important to go through the options in sequence, and we asked them to indicate next to each link whether that was the product for which they were looking (yes or no). When clicked, each link opened a small pop-up window with a picture of a headphones set. All the links except one randomly presented one similar but incorrect target (i.e., a lure), and one link presented the correct target product. In the 2-lure condition, the correct target was the third link in the sequence. In the 6-lure condition, the correct target was the seventh in the sequence.

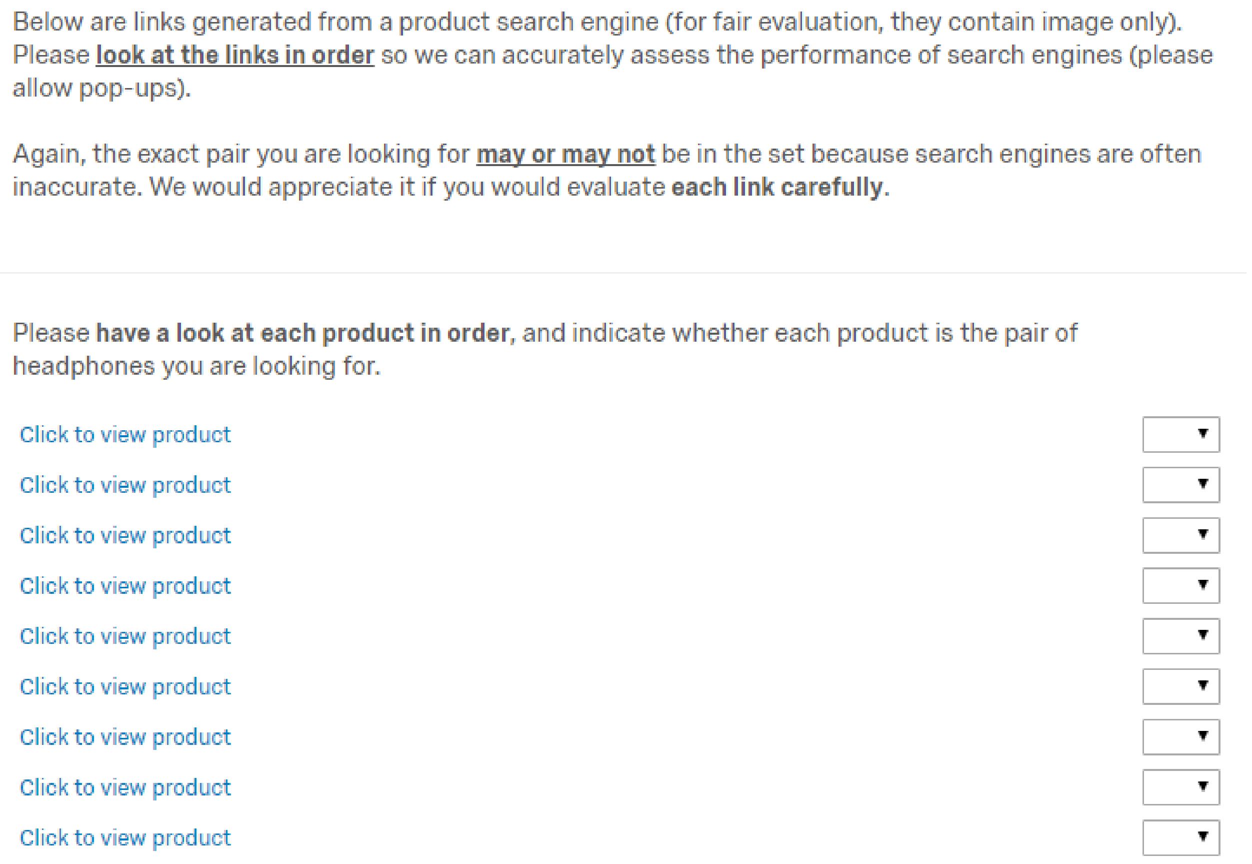
We also manipulated the presence of a distinctive feature in the focal product. In the cue-absent condition, the target was identical to the one used in study 2. In the cue-present condition, the target item had a small logo on one side (see below section). The logo was designed to be noticeable but not obtrusive in order to prevent the recognition task from becoming too easy. Importantly, to rule out the possibility that adding a distinctive feature simply made the identification task easier (and therefore less susceptible to the number of rejected lures in general), we conducted a separate pretest, using the same procedure. We specifically tested whether including a distinctive feature in the focal target made the task easier ("Overall, how difficult was the task to find the product you were looking for?"). Results suggest that task difficulty was not different across conditions (*F*(3, 112) = .29, *p* = .832).

Similar to the previous studies, our focal dependent measure was whether participants correctly identified the true target as a function of its position in the lineup. Note that the links were indistinguishable until clicked, and participants saw eight links regardless of condition (the eighth item listed was a lure in both conditions). In addition to demonstrating generalizability, this design precludes the possibility that base-rate beliefs are driving the effect, because the total number of links visible (i.e., eight) was held constant.

Results

We predicted that the threshold escalation effect to be attenuated when the focal target included a distinctive cue. Confirming our prediction, a linear probability model revealed a significant lure (2 vs. 6) x distinctive cue (present vs. absent) interaction (*t*(367) = 2.40, *p* = .017, *Cohen's d* = .25). A logistic regression model revealed a similar result (*Z* = 2.38, *p* = .017). Specifically, the threshold escalation effect was replicated in the cue-absent condition, where participants were less likely to accurately identify the correct target in the 6-lure (58.5%) than in the 2-lure condition (83.8%; χ^2 (1) = 14.83, p < .001, *Cohen's d* = .41). However, this effect was significantly attenuated in the cue-present condition (6 lures: 78.4% vs. 2 lures: 82.9%; χ^2(1) = .58, p = .45). The difference between the cue-absent and cue-present 6-lure conditions was also significant (cue-absent: 58.5 % vs. cue-present: 78.4%; χ^2 (1) = 8.50, p < .01, *Cohen's d* = .31). Note that these results were not different when we excluded participants who judged more than one item to be the correct target (t(337) = 2.28, p = .023 for interaction; χ^2 (1) = 11.120, p < .001 for cue-absent condition; χ^2 (1) = 0.15, p = .70 for cue-present condition). Ancillary analyses are provided in the last section of this appendix.

2. Product lineup setup used in study 5



Below are links generated from a product search engine (for fair evaluation, they contain image only). Please **look at the links in order** so we can accurately assess the performance of search engines (please allow pop-ups).

Again, the exact pair you are looking for **may or may not** be in the set because search engines are often inaccurate. We would appreciate it if you would evaluate **each link carefully**.

Please **have a look at each product in order**, and indicate whether each product is the pair of headphones you are looking for.

Click to view product ▼

Click to view product ▼

Click to view product ▼

Click to view product ▼

Click to view product ▼

Click to view product ▼

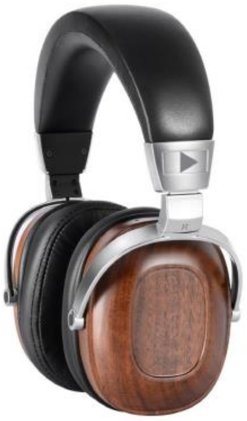
Click to view product ▼

Click to view product ▼

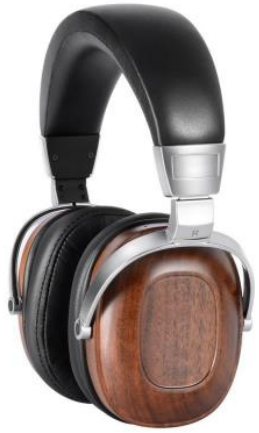
Click to view product ▼

3. Cue manipulation for the correct target in study 5

- Cue present



- Cue absent



4. Ancillary analyses

Alternative accounts. To examine whether the results could be due to increased fatigue, depletion, or confusion in the 6-lure condition, we compared the correct rejection rate of the second judgment in the 2-lure condition with that of the sixth judgment in the 6-lure condition. As in the first follow-up study, the correct rejection rate for these options did not degrade as people were going through a longer sequence, regardless of whether they were given a distinctive cue (2nd judgment: 95.1% vs. 6th judgment: 98.0%) or not (2nd judgment: 97.1% vs. 6th judgment: 96.3%; all *p*'s > .50 respectively, including the two-way interaction). This casts doubt on alternative accounts based on depletion or confusion (see below table for details).

CORRECT IDENTIFICATION RATES IN STUDY 5

Sequence	1st	2nd	3rd	4th	5th	6th	7th	8th
2-lures, cue-absent	94.3%	97.1%	83.8%*	97.1%	96.2%	98.1%	100.0%	95.2%
6-lures, cue-absent	82.9%	89.0%	97.6%	95.1%	95.1%	96.3%	58.5%*	97.6%
2-lures, cue-present	92.7%	95.1%	82.9%*	96.3%	98.8%	97.6%	97.6%	97.6%
6-lures, cue-present	94.1%	95.1%	95.1%	94.1%	98.0%	98.0%	78.4%*	97.1%

asterisk (*) denotes the correct target

Next, we tested whether the results might be driven by self-selection. One might be concerned that those who correctly rejected the six lures positioned before the target (in the 6 lures condition) had a more stringent matching threshold in the first place, compared with those who correctly rejected two lures (in the 2 lures condition). We tested this alternative account using the *matched-choice* paradigm (Gal and Liu 2011), where we compare the identification accuracy for lures presented either before or after the correct target (i.e., in the 6 lures vs. 2 lures condition, respectively). If those who correctly rejected the lures in the 6 lures condition were those who had a stringent threshold to begin with, their identification rate for any given lure, regardless of its position, should be higher than that for those in the 2 lures condition. Casting doubt on this account, a 2 (lure position: before target vs. after target) x 2 (lures: 2 vs. 6) linear probability model with identification accuracy as the dependent variable for each of the 7 lure products revealed no significant main effects nor interactions (all *p*'s >.40), except for one product which showed a marginal interaction (*p* = .091). This suggests that people who correctly rejected six lures before the target were not inherently different, in terms of identification accuracy, from those assigned to the two-lure condition.

d' and C analyses. Two commonly used statistics in signal detection models are *d*' (discrimination sensitivity index) and *C* (criterion index). These measures utilize standardized scores to address the question of whether participants are simply becoming more inclined to respond yes or no (i.e., response bias), or whether identification decisions are truly being influenced (see Swets, Dawes, and Monahan 2000). Thus, the two measures are bias-corrected measures that represent true recognition performance after accounting for a person's tendency to respond in a certain way. The computational formula for standard yes-or-no recognition tasks is:

$$d' = Z_{correct_j} - \mathrm{Z}_{incorrect_k} \qquad (1)$$

Here, $Z_{correct}$ is the z-score of the hit rate for target j (i.e., correctly identifying the target), and $Z_{incorrect}$ is the z-score associated with falsely responding “yes” to lure product k. *C* is conceptually similar, and measures criterion placement where higher numbers indicate more stringent criterion while lower numbers indicate relatively lax criterion position. *C* is formally derived by the following equation (Macmillan and Creelman 2004):

$$\text{(2)} \qquad C = -\frac{z_{correct_j} + \mathrm{Z}_{incorrect_k}}{2}$$

Although we are only interested in recognition of a single target, these measures raise methodological problem as they require multiple targets and multiple lures in order to be computed. To rectify this issue, we follow the previous suggestions (Macmillan and Kaplan 1985; Cradit, Tashchian, and Hofacker 1994; Macmillan and Creelman 2004) and use collapsed *d*' – computed by pooling data across participants by group and then calculating *d*'.

COMPARISON OF INDICES

				95% CI	
	Cue			Lower CL	Upper CL
2-lures	Absent	*d'*	2.849	2.510	3.187
		C	0.438	0.269	0.607
	Present	*d'*	2.765	2.390	3.141
		C	0.431	0.244	0.619
6-lures	**Absent**	***d'***	**1.720**	**1.404**	**2.036**
		C	**0.645**	**0.487**	**0.803**
	Present	*d'*	2.530	2.211	2.850
		C	0.478	0.319	0.638

Note that we use confidence intervals to draw statistical conclusions due to the pooling procedure (Macmillan and Creelman 2004). The results show that participants who saw 6 lures without the presence of additional cue were less likely to correctly identify the true target (*d*' = 1.72, 95% CI [1.40, 2.04]) compared to those in other conditions (see table above for a comparison). More importantly, cue-absent 6-lures condition induced participants to place a significantly higher criterion (*C* = 0.65, 95% CI [0.49, 0.8]) above and beyond those in the other conditions. Indeed, this criterion measure captures both the likelihood of falsely rejecting the correct target and the likelihood of correctly rejecting the false target. However, given the results discussed in the mere error section, the stringent threshold placement seems to be more likely to be driven by higher false rejection of the correct target.

WEB APPENDIX I

Supplementary materials for follow-up of study 5

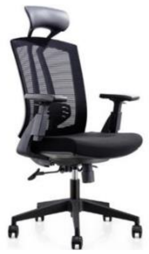
This follow-up study examines how the actual similarity of the lures influences the threshold escalation effect. To do so, we pretested several lure candidates and classified them into either more-similar or less-similar lures, which were then presented to participants in the main study. We also manipulated three different lure quantity conditions to further examine the role of lineup length on our effect.

1. Pretest procedure and results

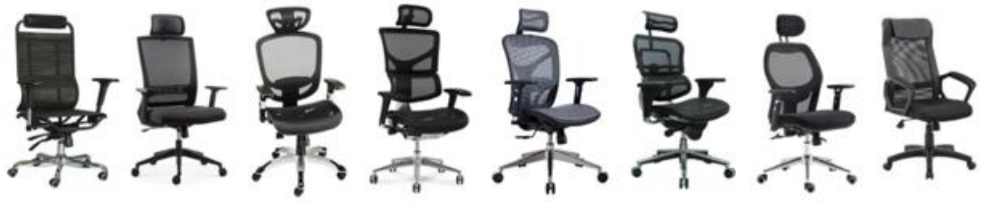
In this pretest, we recruited 80 Master-level participants from MTurk, given the nature of the task. Participants were shown a picture of the correct target product to be used in the follow-up study (i.e., office chair), and were asked to evaluate how similar each of the lure products looked to the correct target (0% – 100%). In total, participants evaluated 16 lure products, each on a single page, in a random order, where 8 of them were classified as more-similar lures and the other 8 as less-similar lures a priori. The pretest was thus conducted to ensure whether the lures selected to be more similar to the target product were actually perceived more similar.

A planned contrast revealed that more-similar lures were judged to be more similar to the true target than less-similar lures ($F(1, 79) = 387.12$, $p < .001$, $\eta_p^2 = .831$). Overall, more-similar lures were considered 67.4% similar, whereas less-similar lures were perceived to be 43% similar to the correct target. Given the strong evidence, we presented either 2, 5, or 8 randomly drawn lures from the set of 8 more-similar (less-similar) lures to those in the high-similarity (low-similarity) condition in the follow-up study. The actual lures and the correct target products presented to participants are shown below.

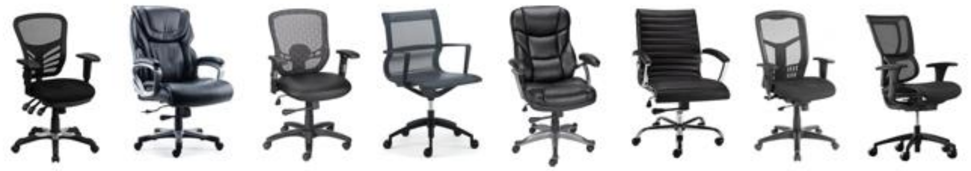
- Correct target:



- More-similar lures:



- Less-similar lures:



2. Follow-up study procedure and results

The follow-up study conceptually replicates the findings of study 4 by manipulating the actual lure similarity, using the stimuli pretested above. In doing so, we also explore the role of lineup length by providing either 2, 5, or 8 lures depending on the condition participants are assigned to, before encountering the correct target. This additional lure sequence condition allows us to examine the frequency of rejection decisions that foster threshold escalation, and whether the relationship between lineup length and misidentification rate is linear or non-linear in nature. Given the exploratory nature of lineup length manipulation, we predicted a main effect of lure similarity, and a main effect of number of lures screened, but not an interaction between the two factors a priori. The current study also uses a different set of products to bolster generalizability of the proposed effect.

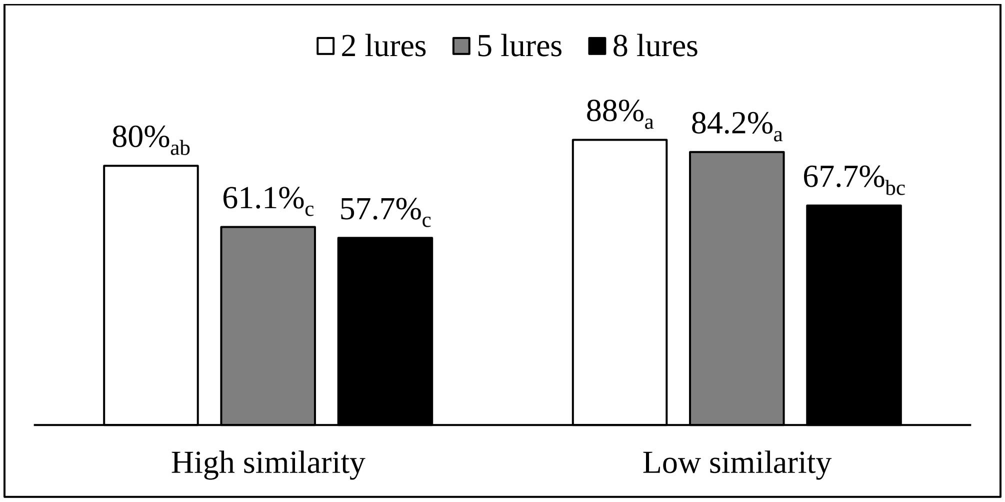
Method

Three hundred and forty participants (M_{age} = 38.7, 51.2% women; excluding 33 participants who did not follow instructions – including them did not significantly alter the results) from MTurk were randomly assigned to one of the following six between-subjects conditions: 2 (similarity: high vs. low) × 3 (lures: 2 vs. 5 vs. 8). All participants saw a picture of an office chair, and imagined that they had a chance to sit on the chair recently. We further told the participants to assume that they were looking for the exact chair as it was that specific product that felt comfortable. After the image of the product disappeared, participants completed the same filler task used in the previous studies, and simulated screening through either 2, 5, or 8 lures before the true target was presented again. The focal dependent variable was whether participants correctly identified the true target. We also measured the amount of time participants spent on the identification task for the correct target.

Those in the high similarity condition were presented lure products which were considered to be more similar to the true target in a separate pretest (see above), while those in the low similarity condition were presented with lures that were considered less so. The order in which the lures were presented were completely randomized.

Results

The results are summarized in the below figure. The interaction between the two factors were not significant ($\chi^2(2) = 1.79$, $p = .41$), and we thus focus on the main effects. Replicating the threshold escalation effect, the main effect of number of lures screened was significant ($\chi^2(2) = 12.58$, $p = .002$, *Cohen's d* = .40). Pair-wise contrasts revealed that participants in 2 lures condition were significantly less likely to falsely reject the correct target than those in the 8 lures condition (2 lures: 83.5% vs. 8 lures: 63.2%; $p < .01$), and marginally less than those in the 5 lures condition (2 lures: 83.5% vs. 5 lures: 73%; $p = .072$). The difference in correct identification rate between 5 lures condition and 8 lures condition were also marginally significant (5 lures: 73% vs. 8 lures: 63.2%; $p = .069$). The main effect of lure similarity was also significant, such that those who screened through more-similar lures were more likely to falsely reject the correct target (high similarity: 67.3% vs. low similarity: 79.3%; $\chi^2(1) = 7.88$, $p = .005$, *Cohen's d* = .31).



**Percentages with different subscripts indicate significant statistical difference p < .05*

Although the interaction between similarity and lineup length was not significant, the pattern of results appears to suggest that when similarity between the lures and the target is high, recognition accuracy decreases significantly already after seeing five lures (61.1%), compared with the two lures condition (80.0%; $\chi^2(1) = 5.16$, $p = .023$). When similarity is low, however (i.e., when the lure is more easily discernible), recognition accuracy appears to decrease only after seeing eight lures (67.7%), compared with five lures (84.2%; $\chi^2(1) = 5.16$, $p = .037$), with no difference between two and five lures (84.2% vs. 88.0%; ns). This preliminary finding warrants further investigation.

Another observation is that identification rates across the sequence of lures in each of the lure similarity conditions appear to follow non-linear patterns. Specifically, in the high-similarity condition, model fit improved when a log function was used (AIC = 212.60), compared with a linear function (AIC = 213.36), consistent with the visibly convex pattern of the results. In the low-similarity condition, model fit improved when an exponential function was used (AIC = 224.87), compared with a linear function (AIC = 225.53), in line with the concave pattern of results. These results provide preliminary evidence that the relationship between the number of lures screened and false rejection effect may be non-linear.

With regards to the time spent on identification decision, no main effect emerged (number of lures: $\chi^2(2) = 2.6$, $p = .272$; lure similarity: $\chi^2(1) =.55$, $p = .458$), and the interaction between the number and the similarity of the lures was also not significant ($\chi^2(2) = 1.26$, $p = .532$). Overall, participants spent on average 10.94 seconds on deciding whether the product was the correct target or not.

Discussion

The results show that the proposed effect is dependent not only on the number of lures, but also on how similar the lures encountered by consumers are to the true target. That is, when the lures are relatively less similar to the target, it seems that a longer sequence of lures is necessary for the threshold to escalate, and therefore increase the likelihood of falsely rejecting the true target. The results further suggest that people's internal matching threshold may not necessarily elevate linearly with the number of lures screened.

The pattern of results is also consistent with the findings in study 5. That is, considering that the presence of a distinctive cue is an extreme version of dissimilarity manipulation, participants in the follow-up study were also less likely to falsely reject the correct target when the lures were less similar to the target product, as those in the cue-present condition of study 5. Indeed, those who screened through less-similar lures also misidentified the correct target after seeing 8 lures, and future research may investigate how longer sequence of lineups may affect identification accuracy.

We also measured the amount of time participants spent on the identification decision and found no difference across conditions. This pattern of results cast doubt on the alternative explanation that encountering more lures or more similar lures simply made the task more difficult or confusing and therefore led to a reduced recognition accuracy. It also seems unlikely that participants were making haste absent-minded guesses given the average time they spent on the decision.

WEB APPENDIX J

Summary of results

CORRECT IDENTIFICATION RATES ACROSS STUDIES

		Fewer lures	More lures	*p*-value
Study 1		65.4%	39.1%	0.009
Study 1 follow-up		82.1%	61.3%	0.004
Study 2		81.1%	55.5%	<.001
Study 2 follow-up 1		77.5%	59.8%	0.019
Study 2 follow-up 2		83.2%	73.5%	<.001
Study 3	control	76.5%	61.4%	0.037
	lax-cue	73.2%	78.4%	0.45
Study 4	baseline	78.7%	55.9%	0.008
	high-difficulty	62.1%	60.0%	0.82
Study 5	cue-absent	83.8%	58.5%	<0.001
	cue-present	81.9%	78.4%	0.45
Study 5 follow-up	low-similarity	88.0%	67.7%	0.007
	high-similarity	80.0%	57.7%	0.008

WEB APPENDIX K

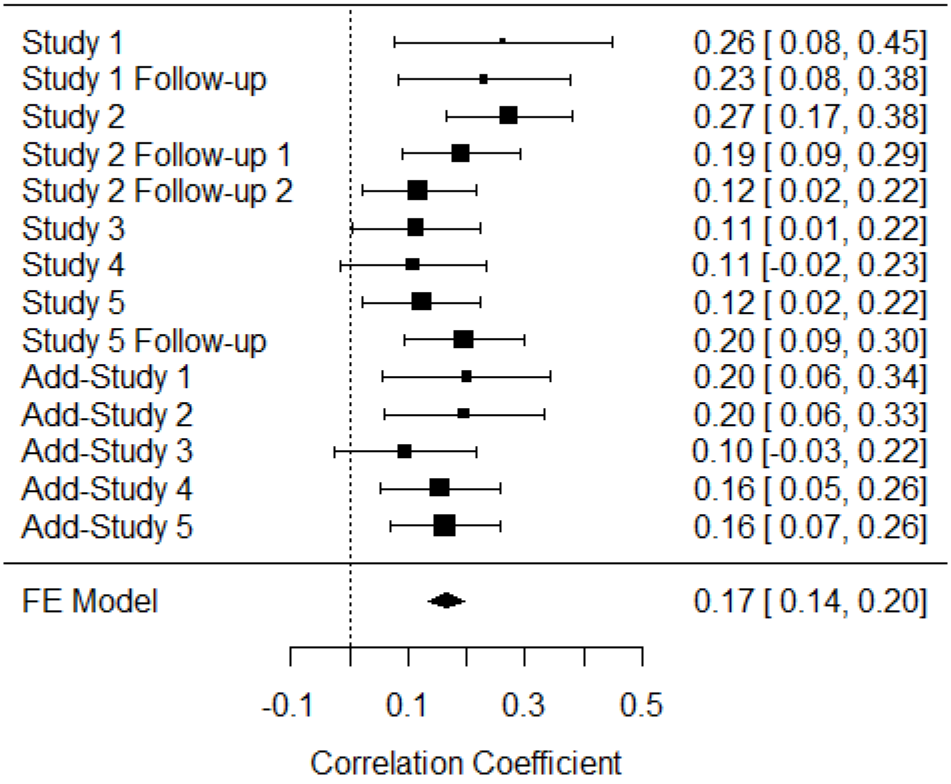
Meta-Analysis

We provide a meta-analysis given its usefulness for testing an overall existence and strength of a novel effect. We include all six studies, three follow-up studies (excluding the follow-up study investigating doubt), and five additional studies not reported here given mediocre incremental contribution.

Additional study 1 (N = 172, MTurk) directly replicates the follow-up of study 1 without providing participants about how many similar products exist in the market (i.e., without the base-rate information). Additional studies 2 (N = 194, MTurk) and 3 (N = 257, MTurk) replicate the proposed threshold escalation effect that screening through a greater number of lures (6 vs. 2) increase the likelihood of falsely rejecting the correct target with different filler questions and with different stimuli (sunglasses), respectively. Additional study 4 (N = 351, MTurk) tests the robustness of the effect by including a condition where participants encounters a product different from the correct target (i.e., another lure) instead of the true target at the end of the lineup, resulting in a 2 (9 vs. 5 lures) x 2 (correct target vs. incorrect target). In line with our prediction, rejecting a greater number of lures induced erroneous judgment (i.e., false rejection) only for the correct target. Additional study 5 (N = 421, MTurk) replicates the results of additional study 4 using simulated rejection paradigm. Details of the additional studies are available from the authors upon request.

We used the ‘*metafor’* package available on R for analysis (https://cran.r-project.org/web/packages/metafor/index.html). In order to account for interaction effects from linear probability models and to avoid potential caveats, we utilized correlation coefficient effect size (*r*) for each study and adopted a simple fixed-effects model for the meta-analysis (Lipsey and Wilson 2001). We included the predicted effects (main effects or interaction effects) for each study. Each study was also weighted by their respective number of participants.

The meta-analysis of the studies estimated an average effect size of *r* = .167 (*Z* = 10.78, *p* < .001). To ease interpretation, we converted this to Cohen's *d* (*d* = 0.34), indicating a small to intermediate effect size. The results clearly suggest that the proposed effect of threshold escalation has sound evidence. Screening a greater number of *innocent*, lookalike products significantly increases the chances that the correct, *guilty* target is falsely rejected. We provide an overview of the results in the figure below. The estimated effect size was not substantially different when the additional studies were excluded from the analysis (*r* = .171, *d* = 0.35).



Note: main effects for studies 1, 2, 5, additional studies 1 – 3 and all follow-up studies were included; interaction effects for studies 3, 4, 6 and additional studies 4 and 5 were included in the analysis.

REFERENCES

Benjamin, Aaron S., and Sameer Bawa (2004), "Distractor Plausibility and Criterion Placement in Recognition," *Journal of Memory and Language*, 51(2), 159-172.

Cradit, J. Dennis, Armen Tashchian, and Charles F. Hofacker (1994), "Signal Detection Theory and Single Observation Designs: Methods and Indices for Advertising Recognition Testing," *Journal of Marketing Research*, 31(1), 117-127.

Lipsey, Mark W., and David B. Wilson (2001), “Practical Meta-Analysis,” *Thousand Oaks*.

Macmillan, Neil A., and C. Douglas Creelman (2004), “Detection Theory: A User's Guide,” *Psychology Press*.

Macmillan, Neil A., and Howard L. Kaplan (1985), "Detection Theory Analysis of Group Data: Estimating Sensitivity From Average Hit and False-Alarm Rates," *Psychological Bulletin*, 98(1), 185-199.

Sela, Aner, and Baba Shiv (2009), "Unraveling Priming: When Does The Same Prime Activate a Goal Versus a Trait?," *Journal of Consumer Research*, 36(3), 418-433.

Swets, John A., Robyn M. Dawes, and John Monahan (2000), "Psychological Science Can Improve Diagnostic Decisions," *Psychological science in the public interest*, 1(1), 1-26.