Journal of Applied Research in Memory and Cognition

Which Energy Label Did That Appliance Have Again? A Memory Test Reveals Confusing Ecolabel Design

Emil Skog, John E. Marsh, and Patrik Sörqvist

Online First Publication, June 15, 2026. https://dx.doi.org/10.1037/mac0000284

AMERICAN PSYCHOLOGICAL ASSOCIATION SARMAC Journal of Applied Research in Memory and Cognition

ISSN: 2211-3681
https://doi.org/10.1037/mac0000284

EMPIRICAL ARTICLE

Which Energy Label Did That Appliance Have Again? A Memory Test Reveals Confusing Ecolabel Design

Emil Skog[1], John E. Marsh[1, 2, 3], and Patrik Sörqvist[1, 4]
[1] Department of Health, Learning and Technology, Luleå University of Technology, Sweden
[2] Human Factors Laboratory, School of Psychology and Computer Sciences, University of Lancashire, United Kingdom
[3] Faculty of Society and Design, Bond University, Australia
[4] Department of Building Engineering, Energy Systems, and Sustainability Science, University of Gävle, Sweden

Mistakenly remembering an energy inefficient appliance as efficient and vice versa can have undesirable consequences. Within the European Union, household appliances (e.g., refrigerators) are assigned energy labels that indicate energy efficiency. Recently, the European Union moved from an old scale (with energy labels A+++, A++, A+, A, B, C, D) to a new scale (A, B, C, D, E, F, G). We investigate how these energy labels are processed from a memory perspective. The results of two experiments found that the "A+ classes" in the old scale were all seen as environmentally friendly. This similarity was associated with worse recall fidelity at a later memory test, suggesting poor distinctiveness in memory. The old scale produced lower recall accuracy and larger confusion errors (e.g., a poorly performing product was mistakenly remembered as highly energy efficient). We identify novel psychological consequences of ecolabeling and discuss their implications for marketing and consumer behavior.

General Audience Summary

Mistakenly remembering an energy inefficient appliance as highly efficient and vice versa can, for example, lead to surprisingly high energy bills, poorly informed purchase decisions, and unnecessary replacement of energy efficient appliances. An ecolabel design that facilitates memory is therefore to be preferred. Within the European Union, household appliances (e.g., refrigerators, freezers) are assigned energy labels that indicate energy efficiency. These labels are used in advertisements and could influence memory in different ways. Recently, the European Union moved from an old scale (with energy labels A+++, A++, A+, A, B, C, D) to a new scale (A, B, C, D, E, F, G). We investigate how these energy labels influence the memory of appliances' energy efficiency. In two experiments, participants were first asked to rate the environmental friendliness of appliances with different energy labels and to later recall which label belonged to which appliance. The results found that the top energy classes of the old scale (the "A+ classes") were all perceived as environmentally friendly, indicating similarity in the way they are understood and processed in memory. This similarity was associated with worse recall fidelity at the subsequent memory test. The energy classes of the old scale were more difficult to accurately recall, and they were also mixed up with other energy classes to a larger degree (e.g., a poorly performing product was mistakenly remembered as highly energy efficient). We discuss the consequences of this novel psychological finding and its implications for consumer behavior, in particular, the design of ecolabels used in marketing.

Keywords: environmental psychology, energy efficiency, ecodesign, source memory, distinctiveness

Supplemental materials: https://doi.org/10.1037/mac0000284.supp

Kamala London served as action editor.
Emil Skog https://orcid.org/0000-0003-2492-9933
John E. Marsh https://orcid.org/0000-0002-9494-1287
Patrik Sörqvist https://orcid.org/0000-0002-7584-2275
The authors declare there are no conflicts of interest, financial, personal, or otherwise. Emil Skog was supported by a grant from Kempestiftelserna (Grant JCSMK23-0179). Patrik Sörqvist was supported by a grant from Stiftelsen Riksbankens Jubileumsfond (Grant P23-0067).

Emil Skog: conceptualization, methodology, software, formal analysis, investigation, resources, data curation, visualization, writing–original draft,

continued

Consumers around the world use information from ecolabels when forming attitudes about products (Akroush et al., 2019; Andor et al., 2020; Davis & Metcalf, 2016; Gorissen & Weijters, 2016; Grankvist et al., 2004; Holmgren et al., 2018; Newell & Siikamäki, 2014; Sammer & Wüstenhagen, 2006; Skourtos et al., 2021; Sörqvist et al., 2025; Stadelmann & Schubert, 2018; Thøgersen et al., 2024). Energy labels are a type of ecolabel that helps consumers evaluate the energy efficiency of electronic products and appliances (Basiru et al., 2024; Wang et al., 2019). Within the European Union (EU), electronic products must be labeled in accordance with the "Ecodesign and Energy Labeling Directive (2009/125/EC) and Regulation (2017/1369)." This labeling system has generated major resource savings since its introduction in 1994 (Ecodesign Impact Accounting Overview Report, 2024). The present study contributes to research aiming to identify ecolabel designs that improve consumer understanding by applying a cognitive perspective to evaluate how consumers process and retain information from different energy labels.

Of particular interest is the ability to accurately remember how energy-efficient appliances are. Misremembering an energy inefficient appliance as efficient, and vice versa, can lead to poorly informed purchase decisions, inefficient appliance use, or unnecessary replacement of efficient appliances. Such memory errors may also undermine household energy literacy over time (cf. van den Broek, 2019). Ecolabel designs that support memory are therefore preferable.

The EU recently changed their scale for labeling the energy efficiency of electronic appliances. There are seven energy classes in a scale, which indicate the most to least energy-efficient alternatives. The previous scale (the *old scale*) used energy classes A+++, A++, A+, A, B, C, and D (Figure 1). A key problem with the old scale is that consumers often find it difficult to interpret the difference between the highest energy classes (the "A+ classes"), which can reduce consumer awareness of energy efficiency (Heinzle & Wüstenhagen, 2012). In 2021, the EU revised and rescaled the energy label scale. In this *new scale*, the seven energy classes are A, B, C, D, E, F, and G, eliminating the "+" categories. The adoption of the new scale is gradual, and many types of electronic appliances within the EU market are still labeled with the old scale.

Perception of Energy Labels

The old and the new scales' effects on consumers are debated. The new scale can increase demand for the most energy-efficient appliances (Faure et al., 2021). But the new scale can also reduce purchase intention (Beck & Toulouse, 2023), in part because the old scale tends to make products seem more environmentally friendly than the new scale (Stasiuk & Maison, 2022). The old scale can inflate the perceived environmental friendliness of products, compared to the new scale (Skog & Sörqvist, 2026), particularly for the A+ classes (Heinzle & Wüstenhagen, 2012). This likely occurs because the A+ classes are judged as more similar; products with the A+ labels receive relatively similar ecoratings (e.g., carbon footprint or environmentally friendliness)—everything with an "A" or better tends to be seen as quite environmentally friendly.

In contrast, energy labels in the new scale tend to be rated such that low-to-high energy classes receive more linear low-to-high ratings of environmental friendliness (Skog & Sörqvist, 2026). We consider this desirable, as participants appropriately distinguish the labels as if they represent linear steps on an interval scale of energy efficiency. This supports better consumer awareness of energy efficiency. As a first step in the present study, we therefore expected that the A+ classes in the old scale would again be rated more similarly than other classes.

Cognitively Confusing Ecolabel Design?

The present study investigates the cognitive implications of this similarity. If the A+ classes are more similar, then there is less mental distance between them, meaning their memory traces are more compressed. When participants rate an energy-labeled appliance, this affords an encoding event where the appliance and its associated energy label may be retained. But items processed with greater similarity (lower distinctiveness; Hunt, 2006) are more likely to be confused at recall. We propose that poor distinctiveness affects labels that appear similar; similarity should lead to less precise memories. Distinctiveness is part of many memory models (e.g., the temporal ratio model, Brown et al., 2007; the retrieving effectively from memory model, Shiffrin & Steyvers, 1997; gist frameworks, Bartlett & Kintsch, 1995; and fuzzy trace theory, Brainerd & Reyna, 2002). Distinctiveness affects recall at both encoding and retrieval (Schmidt, 1985; Waddill & McDaniel, 1998). For example, manipulating dissimilarity processing at encoding (Hunt & Smith, 1996; Smith & Hunt, 2000) and manipulating the distinctiveness of retrieval cues (Mäntylä & Nilsson, 1988) modulate later recall. Other related similarity effects have been observed with, for example, phonological stimulus properties (Baddeley, 1966; Conrad, 1964; Logie et al., 2000; Nairne, 1990; Oberauer, 2009; Wickelgren, 1965). Letters that sound similar are more likely confused: the phonological similarity effect. As a possible distinctiveness mechanism, we will examine if phonological similarity among the letters in the energy labels can explain how the labels are recalled.

Our novel approach is to investigate recall fidelity with EU energy labels to identify good and bad ecolabel design from a cognitive perspective. Memory for energy labels should be influenced by the distinctiveness of the labels. Labels interpreted as more similar at an encoding event may occupy a more overlapping or compressed space in memory, affecting all stages of memory and harming recall (Hunt, 2006; Oberauer & Kliegl, 2006).

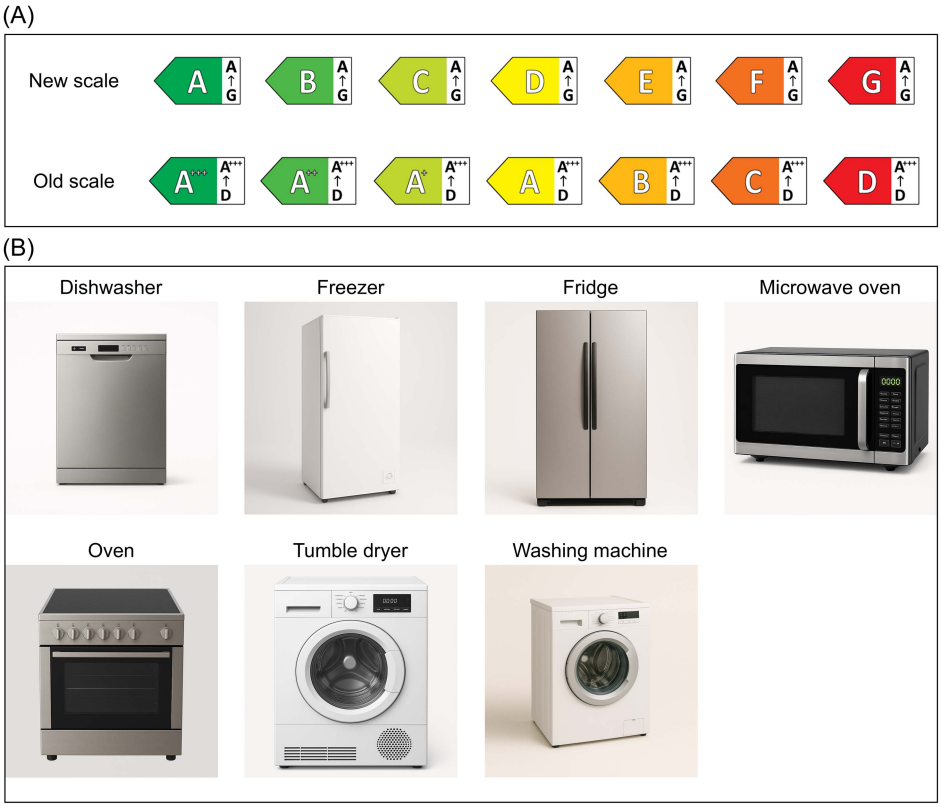
writing–review and editing. John E. Marsh: conceptualization, methodology, formal analysis, writing–review and editing. Patrik Sörqvist: conceptualization, methodology, writing–review and editing, supervision, funding acquisition, project administration.

Emil Skog played a lead role in conceptualization, data curation, investigation, resources, software, validation, visualization, and writing–original draft and an equal role in formal analysis, methodology, and writing–review and editing. John E. Marsh played a supporting role in conceptualization and an equal role in formal analysis, methodology, and writing–review and editing. Patrik Sörqvist played a lead role in funding acquisition, project administration, and supervision, a supporting role in conceptualization, and an equal role in methodology and writing–review and editing.

Correspondence concerning this article should be addressed to Patrik Sörqvist, Department of Health, Learning and Technology, Luleå University of Technology, Laboratorievägen 14, 971 87 Luleå, Sweden. Email: patrik.sorqvist@ltu.se

Figure 1

Energy Labels and Appliances Used as Stimulus Material



Note. Panel A shows the two scales, each containing seven energy labels. Labels from a scale were randomly paired with household appliances (Panel B) to create appliance–label pairs, used as stimulus material. See the online article for the color version of this figure.

Study Overview

In two experiments, we present results that reveal how the EU's old and new energy labels contribute to environmental friendliness judgments, and how these judgments relate to memory. We demonstrate novel evidence on how the design of EU's energy labels affect memory and cognition. Our main hypothesis in both experiments was that the old scale (particularly the A+ classes) would be seen as more similar, which should lead to worse recall fidelity. We operationalized recall fidelity with two measures: Recall accuracy (percent correct) and confusion distance. To analyze confusions, we constructed confusion matrices that helped to identify patterns in how participants confused energy labels. This analysis has the potential to reveal a confusing ecolabel design. We expected larger confusion errors (i.e., a larger scale distance between the erroneously recalled label and the correct answer) with the A+ classes, because their encoded memory trace should be more compressed due to similarity in memory. If part of the label scale occupies a compressed space in memory, then the mental distance between response options in the scale should be smaller, and confusion errors should be larger at recall. Contrariwise, if memory space is not compressed, then items should be more easily discriminated at recall as there is a larger mental distance between the response options. Thus, since parts of the old scale (A+++ to D) are more compressed compared to the new scale (A to G), we predict larger confusion errors in recall of energy classes from the old scale.

Our hypothesis required an encoding task followed by a later recall task. In Experiment 1, participants first rated the environmental friendliness of different appliances, paired with energy labels. In Experiment 2, participants conducted three instead of one rating task. These rating tasks afforded encoding of the stimulus material and served as a manipulation check of perceived label similarity. Later, participants completed an incidental source-memory test, recalling which label belonged to which appliance. We used an incidental ("surprise") memory task instead of explicitly asking participants to be prepared to recall the stimuli because consumers rarely attempt to recall labels that they see and evaluate. In everyday scenarios, information is often encoded incidentally rather than intentionally; an incidental memory test can therefore have greater ecological validity for the study of memory in consumer judgments. The only methodological difference between experiments was that Experiment 1 used one rating task, whereas Experiment 2 used three, repeatedly with the same to-be-recalled material. Since three rating tasks involved repeated encoding that improves recall, Experiment 1 presents results under weaker memory conditions, and Experiment 2 under stronger ones.

Experiment 1

Method

Participants

Participants were recruited via Prolific (https://www.prolific.com) in this internet-based experiment. We conducted an a priori power analysis in G*Power (Faul et al., 2007) to estimate the required sample size to detect the hypothesized difference in confusions across the upper (more compressed) versus lower (less compressed) parts of the old scale. We assumed that this would have a medium effect size of Cohen's $d_z = 0.5$. This estimate of effect size can be considered generic, as we had no prior evidence on recall or confusions to base our estimate on. With a two-tailed hypothesis, $\alpha = .05$, and power ($1 - \beta$ error probability) = .95, the required sample size was estimated to be 54. Since this experiment evaluated two scales, we matched the 54 participants in each of the scale conditions, for a total $N = 108$ (54 female, 54 male, mean age = 46.7 years, $SD = 12.6$). No participant was excluded from analysis. The study conformed to the guidelines of the Swedish Research Ethics Authority (Dnr 2024-05795-01) and the Declaration of Helsinki. All participants were based in the United Kingdom and had English as their first language. They reported having no color-blindness and normal or corrected-to-normal vision. Participants were compensated with £0.75 for a total study time of approximately 5 min, amounting to a per-hour rate of £9.

Materials

The experiment was created in PsychoPy (Version 2024.2.2; Peirce et al., 2019) and JavaScript and was made available online on Pavlovia. Participants used their own desktop computers, in an uncontrolled environment, to access the web-based experiment via Prolific. Participants used their own mouse and keyboard for responses.

The full stimulus material consisted of seven images of household appliances and 14 energy labels; seven labels in each of two scales (Figure 1). Household appliances were of the white goods type; larger electronic household appliances that are often found in kitchens, bathrooms, or laundry rooms. As seen in Figure 1, there was an image of a dishwasher, a freezer, a refrigerator (colloquially named "fridge" in the experiment), a microwave oven, an oven, a tumble dryer, and a washing machine. These images were created with the assistance of a web-based artificial intelligence service and stored in PNG format. Images of appliances were paired with energy labels of the "class arrows" type, made available by the EU in PNG format on the file exchange tool CIRCABC. Images of appliances were presented in the middle of the screen, scaled to 40% of the participant's monitor's height (1:1 aspect ratio). Above each appliance, there was text which clarified what the image represented (the text stated, e.g., "Dishwasher"; see Figure 1). Energy labels were shown to the immediate center-right of the appliances, with a height corresponding to 10% of the participant's monitor (5:3 aspect ratio). The pairing of appliances with energy labels was random, and this random pairing was done anew for every participant. The order of presentation of these stimulus pairs across trials was also random.

Design and Procedure

Experiment 1 was a mixed 2 (Scale: new, old) × 7 (Energy class: first, second, third, fourth, fifth, sixth, seventh) design. Scale was a between-participants factor, where half of all participants were shown the new scale (A to G), and the other half were shown the old scale (A+++ to D). Energy class was a within-participants factor, where all participants completed trials with all seven energy classes in a given scale (Figure 1A). The experiment had two dependent variables: (a) number of correctly recalled energy labels, and (b) the confusion distance of erroneously recalled energy labels. Confusion distances were calculated by identifying incorrectly recalled labels and scoring them based on how far away they were from the correct answer. For example, confusing energy label A with energy label C is a confusion with a distance of two steps across the label scale; therefore, this confusion obtains a confusion distance value of two. A correct response has a confusion distance of zero. The analysis of confusion distances has the potential to reveal where larger confusion errors occurred.

On Prolific, participants were told that this experiment was about evaluating the environmental impact of electronic household appliances. Participants were also told that further instructions would follow once the experiment had begun. Participants provided informed consent before starting the experiment. There was one instruction screen before the initial rating task. This instruction screen stated that each appliance will have an associated energy label, and that the energy efficiency of an appliance can be measured and assigned a value, ranging from very efficient to very inefficient. A graphical representation of the full scale of seven energy classes was shown, making the seven steps of the scale explicit and revealing the color coding of the labels. Here, text stated that the most efficient value is "A" (or "A+++," depending on condition), shown in green color. Text also stated that the least efficient value is "G" (or "D," depending on condition), shown in red color.

Participants then took part in three experimental phases: (a) an encoding phase, (b) a retention interval, and (c) a recall phase. In the first phase, participants conducted a rating task with seven trials, which used each of the seven appliances and labels. On a trial, participants saw an appliance + label stimulus pair and were asked to rate how environmentally friendly the appliance was. Participants responded by clicking an 11-point slider scale presented at the bottom of the screen. The leftmost slider response option stated "Not environmentally friendly" and the right-most option stated "Very environmentally friendly." These ratings afforded a manipulation check of perceived stimulus similarity during encoding.

Upon completing the rating task, participants were told that they were now entering the second part of the experiment, where their task now was to answer a few math questions with a limited time for responses. This was a filled retention interval task, situated between the encoding and recall phases. Participants pressed a button to begin, and six sequential math questions followed, each presented for 10 s. Math questions were in the form: "Is $(6 \times 3) + 2 = 16$"? Participants answered by pressing the "y" or "n" buttons on their keyboards for yes and no responses, respectively. A progress bar indicated the time remaining, and participants had 10 s to give a response before the next question automatically started. If a response was given before 10 s had passed, the computer still waited for the full 10 s before proceeding to the next question. Responses to this filler task were not analyzed.

Upon finishing the 60 s filled retention interval task, participants were then told that the final part of the experiment was next. Their task was now to recall which energy labels were paired with which appliances. To the participants, this was a surprise memory test, as

no previous indication had been given that memory for the stimuli would be tested. The memory test was a source-memory test, where an appliance (without its energy label) was shown as a cue for recall in the middle of the screen. Below the appliance was a row with all seven energy labels (displayed in random order). Participants were prompted to click the label that they believed was the one that was previously paired with the displayed appliance. Selecting an energy label ended a trial. There were seven trials, one for each appliance, and the order of presentation for the appliances was randomized anew so that the appliances were not presented in the same order as during the earlier rating task. The location and thus spatial order of presentation for the seven clickable energy labels were randomly shuffled on each new trial, to mitigate order biases in responses and biases associated with screen regions.

Upon completing the source-memory test, participants were asked yes/no control questions. Later, these were used to verify that no participant had taken notes of the stimulus materials when these were presented in the rating task and that no participant had restarted the experiment. These behaviors could have given a participant an unfair advantage in the memory test. Finally, participants were shown a debrief outlining the purpose of the experiment, before being taken back to Prolific.

Open Science and Analytic Approach

We declare our hypotheses, methods, analyses (statistical tests conducted in Jamovi, Version 2.3.28.0, and R), and results. Experiment 1 was not preregistered, but Experiment 2 was. The data and preregistration supporting this article are publicly available via the Open Science Framework (https://doi.org/10.17605/OSF.IO/FKYDZ). Recall accuracy was analyzed using both aggregated analyses of variance (ANOVAs), reported for continuity with prior work and for visualization purposes, and trial-level generalized linear mixed-effects models (GLMMs). The GLMMs included crossed random intercepts for participants and appliance images to account for individual differences in baseline memory performance and variability in stimulus memorability. This mixed-effects approach was included to address concerns about item-level variability and the aggregation of trial data.

Results and Discussion

We expected that the upper part of the old scale, where the "A+" classes (A+++, A++, and A+) are found, should be rated more similarly. In turn, if the A+ classes are more similar at stimulus encoding, they may occupy a more compressed or overlapping space in memory, leading to worse recall and greater confusion errors. In contrast, we did not expect such a similarity/indistinctiveness effect in the lower part of the old scale, or anywhere in the new scale.

Rating (Encoding) Task

Our analysis requires us to first evaluate the rating results—how the energy labels conveyed environmental friendliness. This will provide context that can ground the results of the memory test and function as a manipulation check of perceived similarity among certain energy labels. Rating results from Experiment 1 can be seen in Figure 2A. To investigate how different energy labels affected ratings, we used a repeated measures 2 (Scale: new, old) × 7 (Energy class: first, second, third, fourth, fifth, sixth, seventh) ANOVA that was adjusted with a Greenhouse–Geisser correction. There was a main effect of energy class, $F(4.1, 434.35) = 411.7$, $p < .001$, $\eta_p^2 = .795$, showing the expected result that participants based their ratings on the energy labels. A main effect of scale, $F(1, 106) = 108.0$, $p < .001$, $\eta_p^2 = .504$, revealed that participants rated the new and the old scale differently. This effect was further explained by an interaction between Energy Class × Scale, $F(4.1, 434.35) = 59.87$, $p < .001$, $\eta_p^2 = .164$. As can be seen in Figure 2A, the energy classes of the old scale were generally rated higher than the energy classes of the new scale, and this difference was greatest at the intermediate energy classes. Energy classes in the new scale were rated with a more linear trend, where low-to-high energy classes received more linear low-to-high ratings of environmental friendliness. But there was a less linear trend across energy classes in the old scale. While participants clearly differentiated their ratings of the lower energy classes, the upper A+ classes were all seen as quite environmentally friendly, as shown by the smaller difference in ratings among these classes (old scale; Figure 2A). Thus, the rating task confirmed that the top energy classes of the old scale are perceived as more similar than other classes.

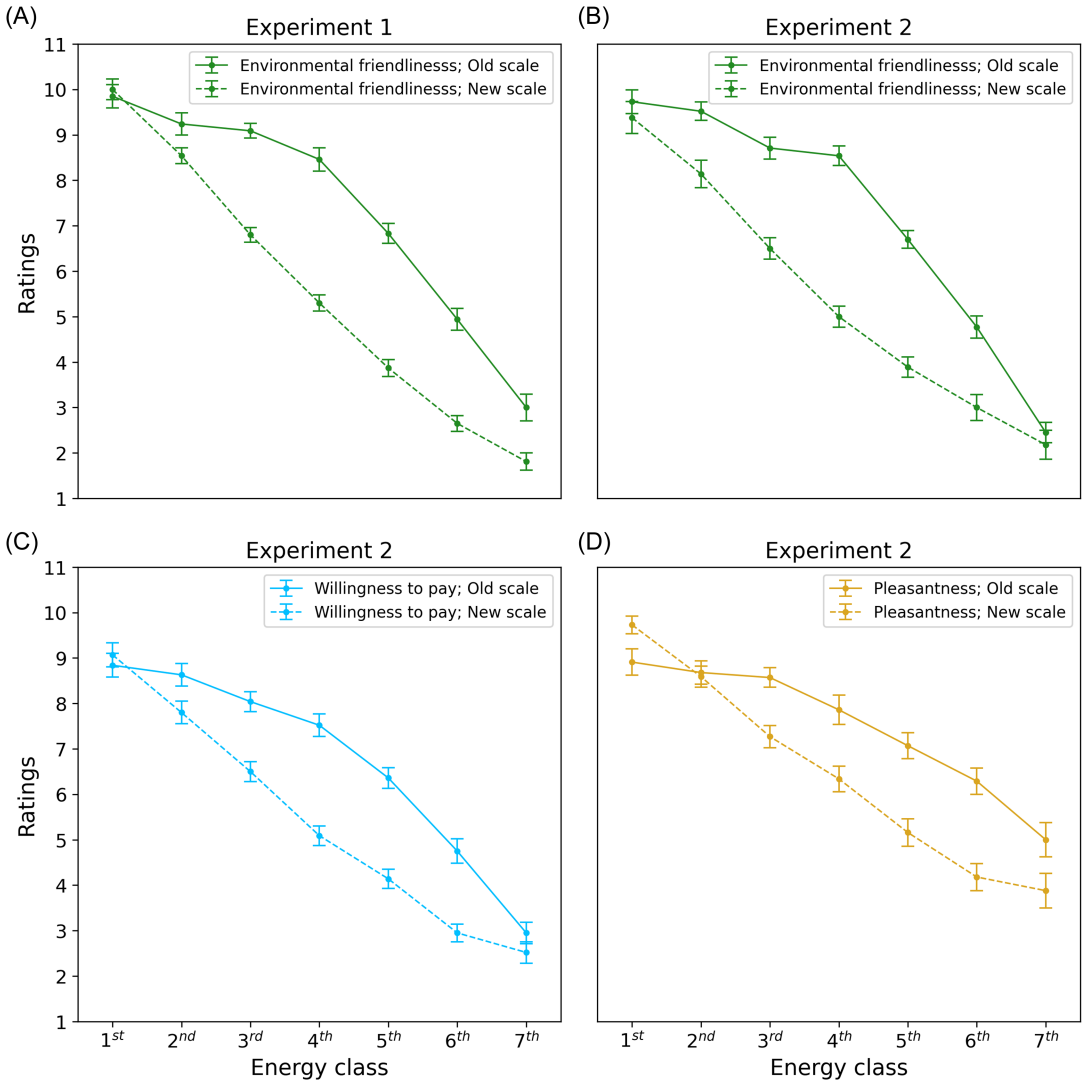
Memory Test

We now turn to our cognitive perspective to examine how processing differences during stimulus encoding might affect subsequent recall and confusion among labels. In the source-memory task, an appliance was shown, and participants were asked to click the energy label they believed was previously paired with the appliance, out of seven options corresponding to the seven possible energy labels in a scale. For each stimulus (appliance + label pair), there was one way to respond correctly and six ways to respond incorrectly. To identify patterns in incorrect responses (i.e., confusions), we constructed confusion matrices for each scale (see the top matrices of Figure 3 for the results of Experiment 1). All analyses of the memory test are based on data shown in these confusion matrices. The main diagonal of each matrix (top-left to bottom-right), where image and response correspond, shows correct responses (recall accuracy). A confusion occurred when a response corresponded to an incorrect image—a deviation from the main diagonal. For our analysis, we will use the two outcomes recall accuracy and confusion distances.

Recall Accuracy. The first performance measurement, recall accuracy, was split by the two scales and the seven energy classes in a repeated measures 2 (Scale: new, old) × 7 (Energy class: first, second, third, fourth, fifth, sixth, seventh) ANOVA. Results are shown in Figure 4A. There was no significant main effect of scale, $F(1, 106) = 1.11$, $p = .294$, $\eta_p^2 = .010$, or interaction between the two factors, $F(6, 636) = 1.87$, $p = .084$, $\eta_p^2 = .017$. The lack of any differences across the scales runs counter to our main hypothesis that memory should be worse for the old scale, particularly the A+ classes. However, the recall accuracy measurement in Experiment 1 was likely undermined by a floor effect where accuracy was not far above chance level. Across all conditions, participants gave an average of 1.78 correct responses, where 1.0 correct responses were chance-level guessing. Though recall accuracy was a limited measurement in Experiment 1, we expected that the second performance measurement, confusion distances, could provide complementary insights, since it is an analysis of patterns in the incorrect responses.

Figure 2

Rating Responses From the Encoding Task, From Both Experiments



Note. Ratings from Experiment 1 (Panel A) and Experiment 2 (Panels B, C, D). Error bars represent the standard errors of the means. See the online article for the color version of this figure.

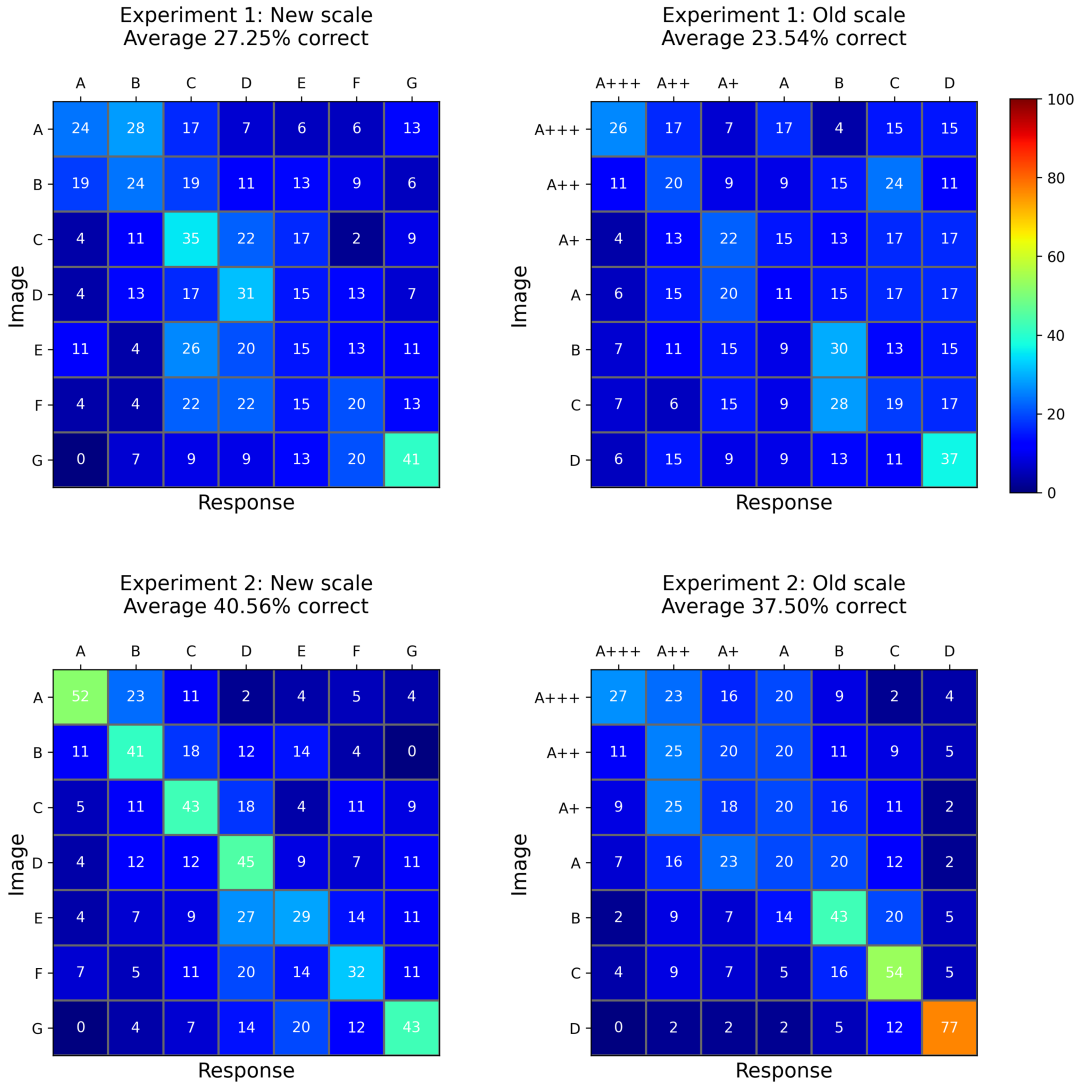
Recall Accuracy: Mixed-Effects Analysis. In the present study, energy labels were randomly paired with appliances, meaning that different participants saw different appliance–label combinations (Figure 1). The ANOVA, therefore, assumes that all appliances were equally memorable, as it only considers the effects of the seven energy classes and the two scales. In a complementary analysis that takes participant-appliance randomization into account by treating these as random effects, we analyzed recall accuracy at the trial level using a GLMM with a binomial error distribution and logit link function. Fixed effects were energy class (first, second, third, fourth, fifth, sixth, seventh), scale (old vs. new), and their interaction. To account for individual differences in overall memory performance and variability in inherent memorability across appliance images, random intercepts were included for participants and appliances.

The model provided limited evidence for an Energy Class × Scale interaction, $\chi^2(6) = 11.48$, $p = .075$ (likelihood-ratio test). Thus, in Experiment 1, the relationship between energy class and source-memory accuracy did not differ reliably across the old versus new labeling scales at conventional significance thresholds. Given the near-floor memory performance observed in Experiment 1, this lack of a reliable interaction is consistent with limited sensitivity to detect scale-related differences in recall accuracy under single-encoding conditions.

To aid interpretation, predicted probabilities of correct source memory were computed from the fitted model (Table 1). Under the old scale, source-memory accuracy varied modestly across energy classes (range ≈ .10–.36). Under the new scale, predicted accuracy was generally similar, although the pattern across classes differed descriptively (e.g., a higher predicted accuracy for Energy Class 4 under the new than old scale).

Figure 3

Confusion Matrices From Both Experiments and Scale Conditions



Note. Confusion matrices contain the accumulated results of all participants. The top and bottom pairs of matrices show the results of Experiments 1 and 2, respectively. Left- and right-hand matrices show results from the new and old scales of energy labels, respectively. Cell values have been converted from total number of responses to percentages, with banker's rounding. See the online article for the color version of this figure.

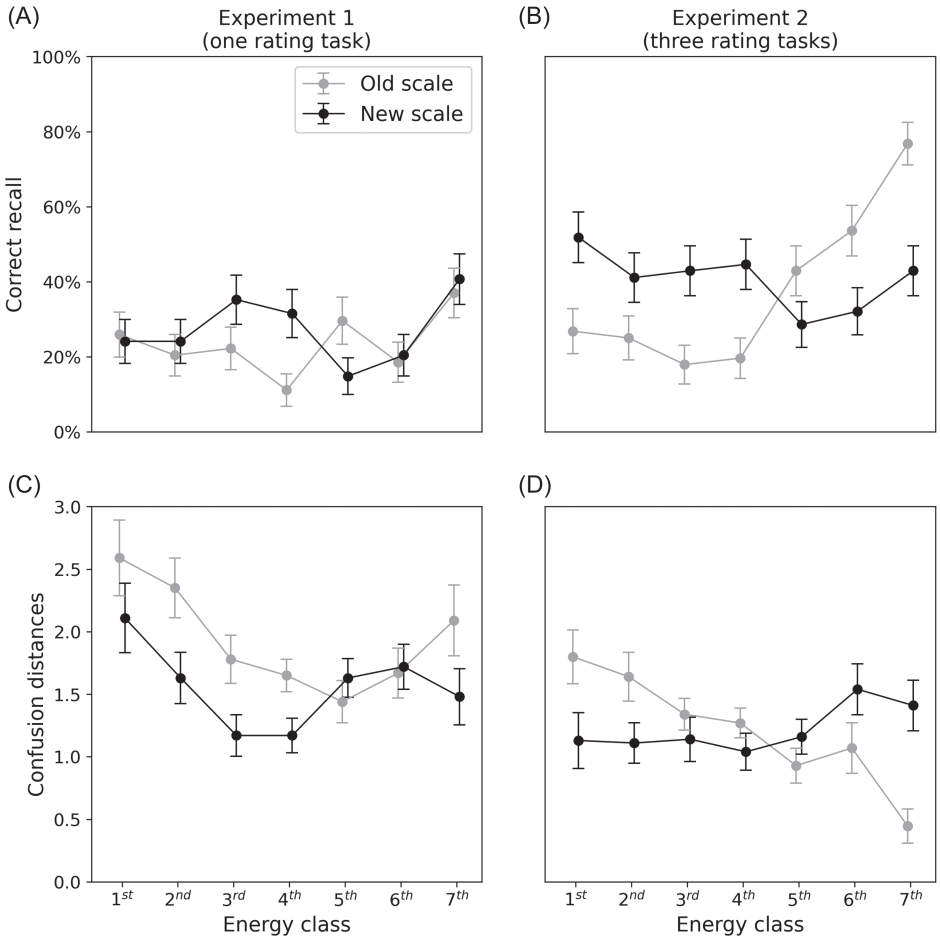
Follow-up comparisons contrasting old versus new labels within each energy class (Holm-adjusted; Supplemental Table S1) showed that Energy Class 4 differed significantly between scales (Old < New), odds ratio = 0.26, $p = .012$. No other energy class showed a reliable old–new difference after adjustment.

Importantly, the mixed-effects analysis indicates that the absence of a reliable scale-related effect in Experiment 1 cannot be attributed to idiosyncratic appliance–label pairings or a small subset of particularly memorable items. Rather, the pattern is consistent with a general lack of discriminative source memory under single-encoding, incidental learning conditions. This supports the interpretation that the hypothesized similarity-based memory disadvantage requires stronger encoding to become detectable, a prediction directly tested in Experiment 2.

Confusion Distances. Confusion distance indexes the magnitude of misremembering on the energy efficiency scale; larger distances correspond to errors that reverse the practical implication of the label (e.g., inefficient remembered as efficient). The confusion distances measurement of memory performance was analyzed with the same 2 × 7 ANOVA with a Huynh-Feldt correction. Results are shown in Figure 4C. Central to our hypothesis, a main effect of scale, $F(1, 106) = 7.77$, $p = .006$, $\eta_p^2 = .068$, revealed that the old scale led to larger confusion errors than the new scale (Figure 4C). Participants did not generally have good memories of the stimuli in Experiment 1. As such, accurate recall was low and confusions were widely spread out. Yet, the significant effect of scale shows that the confusion errors were larger for the old scale than the new scale. An

Figure 4

Recall Accuracy and Confusions Distances From Both Experiments



Note. Recall accuracy in percentages (top panels, A and B) and confusion distances (bottom panels, C and D) from both Experiment 1 (left panels, A and C) and Experiment 2 (right panels, B and D). Error bars represent the standard errors of the means.

example of a common large confusion is that 24% of responses identified the "A++" label as a "C" in the old scale (top-right of Figure 3). This error has a confusion distance value of four. We argue that, from an applied perspective, larger confusions are worse than smaller confusions. A large confusion is an instance where labels are badly misremembered, where "high" and "low" energy efficiency can get mixed up. But a small confusion is an instance where labels are remembered approximately correct, where the selected label has an environmental valence that is close to the correct answer. An assumption in our analysis of confusion distances is that better memory (even on trials with incorrect responses) should lead to smaller confusion errors, as the memory is approximately correct. For example, a participant might remember that the freezer was "somewhat environmentally friendly." Even if the participant does not remember the exact label that the freezer had, they will limit their response to some label that corresponds to their gist memory of environmental valence. Such a gist memory should increase the chances of a correct guess or reduce the size of a confusion error.

We expected that the confusion errors would be worse for the old scale, but also that these problems would be found primarily with the upper part of the old scale (A+ classes), and less so with the other parts of either scale. In the 2 × 7 ANOVA, there was no significant interaction between Scale × Energy Classes, $F(5.22, 553.32) = 1.53$, $p = .175$, $\eta_p^2 = .014$. However, this hypothesis about sections in the scales warrants a more targeted test than the 2 × 7 ANOVA provides. We thus compared the average of the three upper energy classes to the average of the three lower energy classes, across the two scales, in a 2 × 2 ANOVA with Holm–Bonferroni-corrected pairwise comparisons. This revealed a significant interaction, $F(1, 106) = 4.37$, $p = .039$, $\eta_p^2 = .040$. For the new scale, there were similar confusion distances in the upper ($M = 1.64$, 95% CI from estimated marginal means [1.34, 1.93]) and lower parts ($M = 1.61$ [1.38, 1.84]). These did not differ in a pairwise comparison ($p =$ not significant). The lower part of the old scale showed confusion distances that were comparable to those of the new scale ($M = 1.73$ [1.50, 1.97]), but the upper part of the old scale showed larger confusion distances ($M = 2.24$ [1.94, 2.54]). These significantly

Table 1
Predicted Probabilities of Correct Source Memory by Energy Class and Scale in Experiment 1

	Old scale		New scale	
Energy class	Probability	95% CI	Probability	95% CI
1	.24	[0.14, 0.38]	.22	[0.13, 0.36]
2	.19	[0.11, 0.33]	.23	[0.13, 0.37]
3	.21	[0.12, 0.35]	.35	[0.22, 0.49]
4	.10	[0.04, 0.21]	.30	[0.19, 0.45]
5	.28	[0.17, 0.42]	.14	[0.07, 0.27]
6	.17	[0.09, 0.31]	.19	[0.11, 0.33]
7	.36	[0.24, 0.51]	.40	[0.27, 0.55]

Note. Values are estimated marginal probabilities derived from a logistic mixed-effects model with random intercepts for participants and appliances. CI = confidence interval.

differed in a pairwise comparison, $t(53) = 2.67$, $p = .010$, $d_z = 0.363$. This indicates support for our hypothesis that there were larger confusion errors in the upper part of the old scale (the A+ classes; Figure 4C).

Experiment 2

Experiment 1 showed that the A+ classes in the old scale were all seen as environmentally friendly in ratings, whereas the other energy classes were more appropriately discriminated. The old scale subsequently produced larger confusion errors than the new scale, particularly at the A+ classes. Bridging the results of the rating and memory tasks, it appears that similarity harms recall fidelity. This can explain why the more similarly rated A+ classes in the old scale generated larger confusion errors in the source-memory test.

A limitation in Experiment 1 was that source memory was poor, which constrained sensitivity in both the ANOVA and mixed-effects analyses. Floor effects could have masked or attenuated differences that would appear under stronger encoding. In response, we designed a preregistered Experiment 2 where participants encoded the stimulus material three times instead of just once. We added two rating tasks. We expected this to boost recall performance and reduce confusion distances, compared to Experiment 1. Our main hypothesis remained the same: items that appear more similar should be more difficult to recall. Specifically, the A+ classes in the old scale should show evidence of similarity in ratings, which should subsequently lead to worse memory for these A+ energy labels.

Method

Participants

We recruited 112 participants (62 female, 50 male, mean age = 47 years, SD = 14.3) via Prolific who had not previously participated in Experiment 1. We conducted an a priori power analysis to determine the required sample size, based on the effect size of how the confusion distances varied across scales in Experiment 1 ($p = .006$, Cohen's $d = 0.537$). With a two-tailed hypothesis, $\alpha = .05$, and power ($1 - \beta$ error probability) = .80, the required sample size was estimated to be 56 in each of the two scale conditions. No participant was excluded from analysis, as no participant reportedly restarted the experiment or took notes of the stimulus materials during the rating tasks. We adhered to the same participant inclusion criteria and ethical considerations as previously declared under Experiment 1. Participants were compensated with £0.9 for a total study time of approximately 6 min.

Materials, Design, and Procedure

The preregistration of Experiment 2 is available on the Open Science Framework (https://doi.org/10.17605/OSF.IO/FKYDZ). In terms of method, Experiment 2 was like Experiment 1 in all ways except that Experiment 2 introduced two new rating tasks, for a total of three rating tasks. Like in Experiment 1, appliances and labels were randomly paired to create stimuli (Figure 1). The pairs were constant throughout the experiment, but the pairs were presented in a random order in each new task. Participants thus saw and rated the appliance–label stimuli three times each before the memory test. Like in Experiment 1, participants were not forewarned about the different tasks in the experiment.

The first rating task was ratings of environmental friendliness—the same rating task as used in Experiment 1. Pause screens with new instructions between rating tasks gave a clear indication that the rating task instructions were changing. The second rating task was a willingness to pay (WTP) task. In the WTP task, participants were tasked to indicate how willing they would be to pay money for the appliances. To further stimulate engagement with the task, participants were asked to imagine that their own household appliances would soon need to be replaced, and that they were planning to buy new appliances. Responses were given on an 11-point slider scale, like the first rating task. The left- and right-most response options on the scale stated "I would be willing to pay very little" and "I would be willing to pay very much," respectively. The third and final rating task asked the participants to rate how pleasant the appliances would be to have and use in their own home. Responses were again given on an 11-point slider scale, with the left- and right-most response options on the scale stating "Not pleasant to have and use" and "Very pleasant to have and use," respectively.

The three rating tasks were completed in this order by all participants. It is therefore likely that the responses to the second and third rating tasks were influenced by responses to the first rating task. We do not consider this problematic since the results of the WTP and pleasantness ratings were not of interest to our research questions. Further, the ratings of environmental friendliness remained the most important rating outcome (as it functions as a manipulation check of encoding similarity), and with this procedure, it was uncontaminated by the other rating tasks. Inspired by the levels-of-processing paradigm (Cermak & Craik, 1979; Craik & Lockhart, 1972), we chose to use WTP and pleasantness ratings because these are likely to stimulate and require a "deeper" form of semantically rich processing. Deeper processing of stimuli can lead to better recall, compared to more "shallow" processing (e.g., evaluating surface-level features). We considered this suitable for our study on incidental source memory, as we sought to increase recall chances in Experiment 2.

Results and Discussion

Results of Experiment 2 are presented in the same figures, but in different panels, as the results of Experiment 1. After presenting

the results of Experiment 2, we will present the results of a cross-experiment analysis where we evaluated how recall and confusion distances differed across the two experiments.

Rating (Encoding) Tasks

In Experiment 2, the ratings of environmental friendliness (Figure 2B) replicated the pattern of results seen in Experiment 1 (Figure 2A). The results showed that the A+ classes were all seen as quite environmentally friendly, whereas the other energy classes were more distinguished in terms of their environmental valence. All three rating tasks produced significant main effects and interactions (all ps < .001) when analyzed with 2 × 7 ANOVAs. Figure 2 also shows that the outcomes of the three rating tasks varied some (Figure 2B–2D), which could be expected given the task variation. We do not evaluate any hypothesis based on the results of the WTP (Figure 2C) or pleasantness (Figure 2D) ratings, and we reiterate that these ratings were likely influenced by the initial ratings of environmental friendliness (Figure 2B). Overall, the ratings show that participants treated the A+ classes as if they were more similar than other classes.

Memory Test

Recall accuracy was 13.6% percentage points higher in Experiment 2 compared to Experiment 1 ($p < .001$; see cross-experiment analysis below). This indicates that our manipulation to increase the number of encoding events (rating tasks) to three, instead of one, was successful in increasing incidental source memory for the energy labels. The confusion matrices (bottom row of Figure 3) present the full pattern of responses for each image (energy label), accumulated across all participants.

Recall Accuracy. We begin with examining recall accuracy in Experiment 2, using a 2 (Scale: new, old) × 7 (Energy class: first, second, third, fourth, fifth, sixth, seventh) ANOVA. Results are shown in Figure 4B. There was no main effect of scale ($F < 1$), but there was a significant interaction between Scale × Energy Class, $F(6, 660) = 9.45$, $p < .001$, $\eta_p^2 = .079$. As can be seen in Figure 4B, the old and new scales had different patterns of recall accuracy across the energy classes. The new scale showed a relatively stable trend in accuracy across the seven energy classes. But the old scale showed a less stable trend, where accuracy was lower in the upper classes (A+++, A++, A+) and higher in the lower classes (B, C, D). This was supported by a 2 (average of the upper three vs. lower three energy classes) × 2 (old scale vs. new scale) ANOVA, which found a significant interaction, $F(1, 110) = 41.10$, $p < .001$, $\eta_p^2 = .272$. Holm–Bonferroni-corrected pairwise comparisons (Wilcoxon rank) found that recall accuracy was lower in the upper compared to the lower parts of the old scale, $W(55) = 48$, $p < .001$, rank biserial correlation = .877. But there was no such difference within the new scale (p = not significant). This offers strong support for our main hypothesis that memory for the A+ classes should be worse, owing to them being seen as more similar.

Beyond our a priori expectations was the fact that the lowest class in the old scale (class "D") was the class with the highest recall accuracy. Seventy-seven percent of participants accurately recalled this class (Figures 3 and 4B). We consider this congruent with our hypothesis, because if the A+ classes were poorly remembered due to similarity, then the other classes could have become more salient in memory. Participants may have experienced the old scale as having many environmentally friendly labels, which in contrast could have made the least environmentally friendly option more salient.

Recall Accuracy: Mixed-Effects Analysis. Similar to Experiment 1, we analyzed recall accuracy at the trial level using a GLMM with a binomial error distribution and logit link function. Fixed effects were energy class (first, second, third, fourth, fifth, sixth, seventh), scale (old vs. new), and their interaction. To account for individual differences in overall memory performance and variability in inherent memorability across appliance images, random intercepts were included for participants and appliances.

The model revealed a significant Energy Class × Scale interaction, $\chi^2(6) = 56.21$, $p < .001$ (likelihood-ratio test). This shows that our previously discussed scale effects are reliable. The reported source-memory disadvantage for the A+ classes is robust (Figure 4B), and cannot be explained by idiosyncratic pairing with appliances, which may have varied in memorability. To aid further interpretation of this interaction, predicted probabilities of correct source memory were computed from the fitted model (Table 2; Figure 4B). Under the old scale, predicted source-memory accuracy was low for the more efficient energy classes (A, A+, A++, A+++) but increased as energy efficiency worsened, with particularly high accuracy for the least efficient energy class (D). In contrast, under the new scale, this was not seen, though the new scale produced some variability in predicted source memory (Table 2).

Follow-up comparisons contrasting old versus new labels within each energy class (Holm-adjusted) showed that the scale change significantly altered source-memory performance for several classes, particularly at the most efficient and least efficient ends of the scale (see Supplemental Table S3). These comparisons corroborate the overall interaction pattern revealed by the omnibus test. Together, these findings indicate that the revised energy label scale alters not only evaluative judgments but also the memorability of energy efficiency information.

Confusion Distances. Confusion distances were analyzed with a 2 × 7 ANOVA, with a Huynh-Feldt correction. Results are shown in Figure 4D. There was no main effect of scale ($F < 1$), but we found a significant interaction between Scale × Energy Class, $F(5.39, 592.41) = 6.20$, $p < .001$, $\eta_p^2 = .053$. As Figure 4D shows, the confusion distances were more stable with the new scale, but less stable with the old scale. In the old scale, the three upper classes (A+

Table 2
Predicted Probabilities of Correct Source Memory by Energy Class and Scale in Experiment 2

	Old scale		New scale	
Energy class	Probability	95% CI	Probability	95% CI
1	.23	[0.12, 0.39]	.53	[0.35, 0.70]
2	.22	[0.12, 0.39]	.39	[0.24, 0.57]
3	.14	[0.07, 0.28]	.40	[0.24, 0.58]
4	.15	[0.07, 0.29]	.44	[0.28, 0.62]
5	.42	[0.26, 0.60]	.25	[0.14, 0.42]
6	.55	[0.37, 0.71]	.28	[0.15, 0.45]
7	.81	[0.65, 0.90]	.42	[0.26, 0.60]

Note. Values are estimated marginal probabilities derived from a logistic mixed-effects model with random intercepts for participants and appliances. CI = confidence interval.

classes) produced larger confusion distances than the three lower classes. This pattern was confirmed by a 2 × 2 ANOVA, which found a significant interaction, $F(1, 110) = 21.50$, $p < .001$, $\eta_p^2 = .163$. Holm–Bonferroni-corrected pairwise comparisons found that confusion distances were larger in the upper compared to the lower parts of the old scale, $t(55) = 5.35$, $p < .001$, $d_z = 0.715$. But there was no such difference within the new scale (p = not significant). This result closely mirrors what was seen with recall accuracy, where the A+ classes in the old scale produced both lower accuracy (Figure 4B) and larger confusion errors (Figure 4D). In contrast, the new scale showed no significant differences across its upper versus lower energy classes. Taken together, the results of the memory test offer strong support for our main hypothesis.

Cross-Experiment Analysis

The cross-experiment analysis has the potential to reveal differences in the results of the memory test when memory was worse (Experiment 1) versus when it was better (Experiment 2). The analysis was based on a 2 (Experiment: Experiment 1, Experiment 2) × 2 (Scale: new, old) × 7 (Energy class: first, second, third, fourth, fifth, sixth, seventh) ANOVA for each of the dependent variables: (a) recall accuracy and (b) confusion distances.

Recall accuracy significantly differed across the experiments, $F(1, 216) = 23.14$, $p < .001$, $\eta_p^2 = .097$. Rating the stimuli three times led to better recall than rating the stimuli once. Experiment 2's average recall percentage was 39%, and Experiment 1's was 25.4%, a difference of 13.6% percentage points (Figures 3, 4A, and 4B). There were no significant two-way interactions with the experiment factor, but there was a significant three-way interaction across Experiment × Scale × Energy Class, $F(6, 1296) = 3.71$, $p = .001$, $\eta_p^2 = .017$. Recounting the results of Experiment 1, there was no significant effect of scale or interaction between Scale × Energy Class. We attributed this to a floor effect in memory performance, which may have masked any differences across the scales. In contrast, Experiment 2 found a significant interaction between Scale × Energy Class. The three-way interaction, comparing these effects across experiments, revealed that the interaction between Scale × Energy Classes depended on recall accuracy being higher than floor (Experiment 2). This reinforces the conclusion drawn from Experiment 2, that recall was worse for the A+ classes of the old scale (Figure 4B).

The same 2 × 2 × 7 ANOVA, with a Huynh-Feldt correction, was also conducted with confusion distances as the dependent variable. Related to memory being better, there was a main effect across the experiments, $F(1, 216) = 35.23$, $p < .001$, $\eta_p^2 = .140$, indicating that confusion distances were smaller with three encoding events compared to one (Figures 3, 4C, and 4D). There was a significant two-way interaction between Experiment × Scale, $F(1, 216) = 4.56$, $p = .034$, $\eta_p^2 = .021$, confirming that the main effect of scale was only significant in Experiment 1 and not Experiment 2, as previously reported. This outcome was further explained by the pattern of results in the significant three-way interaction, $F(5.36, 1158.82) = 2.37$, $p = .034$, $\eta_p^2 = .011$. The three-way interaction showed that the two scales had slightly different outcomes across the energy classes, and that this also depended on memory for the items. Comparing Figures 4C and 4D, there was a general tendency for the A+ classes to produce larger confusion errors, in both experiments. This is a primary outcome that supports our hypothesis. Of secondary interest, there was also a tendency for the lower parts of the old scale to produce smaller confusion errors, but only when memory was better (Experiment 2), which can explain why the three-way interaction was significant.

General Discussion

The results of two experiments found that the A+ classes in the old scale were evaluated as being more similar. The A+ classes were all seen as environmentally friendly. In support of our hypothesis, this similarity was associated with worse recall fidelity at a later memory test, as the A+ classes produced lower percent correct responses at recall and larger confusion errors.

Memory and Cognition With Different Energy Labels

The present study reveals new details on how different energy labels are cognitively processed. The top (A+) classes of the old energy scale were rated and thus encoded more similarly, even though the color-coding of the energy classes of the old and new scale is identical (see also Skog & Sörqvist, 2026). A set of items that are more similar in memory may occupy a more compressed or overlapping psychological space (Hunt, 2006; Hunt & Smith, 1996; Nairne, 1990; Oberauer & Kliegl, 2006). If part of a scale shows mental compression (e.g., the A+ classes in the old scale), then the mental distance between all classes ("A+++" to "D") should be smaller. This predicts that confusion errors would be larger, which was observed with the old scale, particularly at the A+ classes. In contrast, energy labels in the new scale did not show the same tendency for similarity. The new scale produced relatively smaller (Experiment 1) and more stable (Experiment 2) confusion errors, which indicate that energy classes "A" to "G" are less compressed (more distinct) in memory.

A full examination of exactly why the A+ classes were treated more similarly in the rating task is beyond the scope of the present study. But we argue that the results cannot be straightforwardly explained by phonological similarity (Baddeley, 1966; Conrad, 1964), despite the energy labels being based on letters. The scales contain multiple letters that are of a similar phonological character: Letters "b," "c," "d," "e," and "g" all have a high degree of phonological similarity in English. But letters "a" and "f" are phonologically dissimilar from the rest. If the phonological similarity effect could explain the confusion error results in our experiments, we would expect the more similar items ("b," "c," "d," "e," and "g") to produce larger confusions than the dissimilar items (we especially consider the letter "f" here, since the letter "a" may be confounded by the fact that it is the highest class [new scale] or that there are multiple labels that use this letter [old scale]). But there was no indication that the phonologically dissimilar letter "f" was associated with any higher recall accuracy or smaller confusion error (Figure 4; the "F" class was the sixth energy class in the new scale). Furthermore, we do not consider it likely that the four classes which began with the letter A in the old scale (A+++, A++, A+, and A) were processed with much phonological similarity, given that each added "+" sign should change or add length to a verbal-acoustic representation of the energy class. For example, there should be a large difference in phonological representation between "a" and "a plus." If phonological similarity could explain the old scale's results, then classes A and A+ (phonologically dissimilar) should have been less confusable than classes B, C, and D (phonologically similar). However, the opposite was seen (Figure 3; see especially the bottom-

right confusion matrix, as Experiment 2 provided more robust memory effects).

If not phonological similarity, then what explains the impaired recall for A+ classes? One possibility is visual similarity. Four classes in the old scale have the same letter ("A"), with the only difference being the number of pluses. However, they differ in color-coding, and color is a dominant label feature (Bengart & Vogt, 2023) that tends to grab attention (Lurie & Mason, 2007). Moreover, both the new and old scales were color-coded with a green-red "traffic light" spectrum, but there were still large differences across scales, indicating that color-coding might be a weaker cue than the letters (see Skog & Sörqvist, 2026, for a detailed analysis of how observers assign weight to different label features).

Another possibility is that the negative effect on memory is driven by conceptual similarity among the "A" classes. If participants applied an assumption that a letter scale should begin with the alphabet, then "A-rated" energy efficiency would be representative of very good environmental friendliness. This representativeness heuristic (Tversky & Kahneman, 1974) could explain why every label that contained an "A" was awarded a high rating. This creates similarity among the A and A+ classes on a conceptual level. Further, memory could have been impaired for the A+ classes if participants were less sensitive to the addition of "+" signs. Adding these special characters may have been a less distinct and less processed feature than other features of the labels (e.g., letters or color-coding). In sum, if participants mainly based their ratings on a conceptual understanding of the letters, and were less sensitive to other label features (cf. Skog & Sörqvist, 2026), then classes which contain the letter "A" would be less distinct from each other in memory, and their retrieval would be more based on gist ("high environmental friendliness") than exact energy label (e.g., Brainerd & Reyna, 2002). Consistent with this, the results of the memory test show that the retrieval of the A+ classes was less successful and more confused.

We have discussed how encoding tasks (ratings) influence later recall. But the observed indistinctiveness of the A+ classes is not isolated to the encoding stage of memory. Instead, indistinctiveness or compression of mental space affects all stages of memory (i.e., also storage and retrieval). Other theoretical frameworks on memory support our distinctiveness interpretation. According to the Bartlett and Kintsch (1995) framework of gist in memory, the less distinct A+ classes might activate a schema of "environmentally friendly" (like the representativeness heuristic). This predicts confusion as exact labels within the gist representation are more difficult to accurately retrieve. Similarly, fuzzy trace theory can explain how the A+ classes may have been processed with gist, leading to lower distinctiveness, while other classes using unique letters may have been processed verbatim, leading to higher distinctiveness (Brainerd & Reyna, 2002). Other influential memory models highlight how distinctiveness affects recall (e.g., Brown et al., 2007; Shiffrin & Steyvers, 1997), which helps explain our result that items which appear more similar (A+ classes are all seen as environmentally friendly) are more difficult to recall.

Limitations and Suggestions for Future Studies

Our stimulus set was intentionally small (seven appliances paired with seven labels per scale), and the retention interval was brief. This design offered a tight, tractable test of whether label similarity predicts later confusion, but it also constrains generalizability: real purchase environments contain many product categories and competing visual cues, and memory is often probed after longer delays. We also used an incidental ("surprise") source-memory test to mirror the fact that consumers rarely encode labels with an explicit intention to remember them. However, in everyday life, energy labels may be encountered repeatedly across browsing sessions, advertisements, and in-store comparisons over days or weeks. Future work should therefore examine whether the same confusion patterns emerge under repeated naturalistic exposure and longer retention intervals, ideally using field or longitudinal designs (e.g., repeated browsing tasks, delayed tests after a week, or follow-up memory assessments after real purchases).

Implications for Policy, Label Design, and Consumer Behavior

The present study identifies a cognitively confusing form of ecolabel design. When labels within the same scale are weakly differentiated, they lead to similar perceptions of environmental friendliness (Heinzle & Wüstenhagen, 2012; Skog & Sörqvist, 2026) and, as shown here, also impair memory for energy efficiency information. Specifically, recall fidelity was lower and confusion errors were larger for labels that were psychologically similar, indicating that indistinct category structure carries downstream consequences for memory and decision-relevant cognition. Such memory imprecision may contribute to poorly informed purchase decisions or inefficient appliance use, particularly when consumers rely on remembered rather than currently visible label information.

In contrast, these problems were attenuated when energy classes were more clearly differentiated. The revised EU energy label scale was less prone to recall errors and confusion, suggesting that its categorical structure, using unique letters, affords greater psychological distinctiveness. From a design perspective, this highlights the importance of avoiding within-category variants that preserve the same categorical letter (e.g., A+, A++), as these invite gist-based encoding ("it's A-rated") and reduce discriminability at retrieval. Assigning unique categorical letters to each class appears to be an effective way of supporting distinctiveness and more precise memory representations. Other design features, such as special characters or color-coding, may either exacerbate or mitigate confusion depending on whether they increase or decrease perceived similarity between classes, underscoring the need to evaluate ecolabel features not only in terms of perception and choice, but also memory.

References

Akroush, M. N., Zuriekat, M. I., Al Jabali, H. I., & Asfour, N. A. (2019). Determinants of purchasing intentions of energy-efficient products: The roles of energy awareness and perceived benefits. *International Journal of Energy Sector Management*, *13*(1), 128–148. https://doi.org/10.1108/IJESM-05-2018-0009

Andor, M. A., Gerster, A., & Sommer, S. (2020). Consumer inattention, heuristic thinking and the role of energy labels. *The Energy Journal*, *41*(1), 83–112. https://doi.org/10.5547/01956574.41.1.mand

Baddeley, A. D. (1966). Short-term memory for word sequences as a function of acoustic, semantic and formal similarity. *Quarterly Journal of Experimental Psychology*, *18*(4), 362–365. https://doi.org/10.1080/14640746608400055

Bartlett, F. C., & Kintsch, W. (1995). *Remembering: A study in experimental and social psychology* (2nd ed.). Cambridge University Press. https://doi.org/10.1017/CBO9780511759185

Basiru, I., Xu, Y., Arkorful, V. E., Lugu, B. K., & Ibrahim, A. H. (2024). Energy efficiency labels and urban residents' intention to purchase energy-

efficient household appliances: An empirical study based on the theory of planned behavior. *Clean Technologies and Environmental Policy*, *27*(8), 3433–3448. https://doi.org/10.1007/s10098-024-03070-z

Beck, M., & Toulouse, N. Ö. (2023). Assessing the impact of energy labels on attitude and behavioral intention: An empirical investigation. *Journal of Cleaner Production*, *415*, Article 137751. https://doi.org/10.1016/j.jclepro.2023.137751

Bengart, P., & Vogt, B. (2023). Effects and interactions of labels' color scheme and the individual difference variable lay rationalism on pro-environmental choices. *Journal of Environmental Psychology*, *87*, Article 101998. https://doi.org/10.1016/j.jenvp.2023.101998

Brainerd, C. J., & Reyna, V. F. (2002). Fuzzy-trace theory and false memory. *Current Directions in Psychological Science*, *11*(5), 164–169. https://doi.org/10.1111/1467-8721.00192

Brown, G. D. A., Neath, I., & Chater, N. (2007). A temporal ratio model of memory. *Psychological Review*, *114*(3), 539–576. https://doi.org/10.1037/0033-295x.114.3.539

Cermak, L. S., & Craik, F. I. (1979). *Levels of processing in human memory*. Lawrence Erlbaum. https://doi.org/10.4324/9781315796192

Conrad, R. (1964). Acoustic confusions in immediate memory. *British Journal of Psychology*, *55*(1), 75–84. https://doi.org/10.1111/j.2044-8295.1964.tb00899.x

Craik, F. I. M., & Lockhart, R. S. (1972). Levels of processing: A framework for memory research. *Journal of Verbal Learning and Verbal Behavior*, *11*(6), 671–684. https://doi.org/10.1016/S0022-5371(72)80001-X

Davis, L. W., & Metcalf, G. E. (2016). Does better information lead to better choices? Evidence from energy-efficiency labels. *Journal of the Association of Environmental and Resource Economists*, *3*(3), 589–625. https://doi.org/10.1086/686252

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, *39*(2), 175–191. https://doi.org/10.3758/BF03193146

Faure, C., Guetlein, M.-C., & Schleich, J. (2021). Effects of rescaling the EU energy label on household preferences for top-rated appliances. *Energy Policy*, *156*, Article 112439. https://doi.org/10.1016/j.enpol.2021.112439

Gorissen, K., & Weijters, B. (2016). The negative footprint illusion: Perceptual bias in sustainable food consumption. *Journal of Environmental Psychology*, *45*, 50–65. https://doi.org/10.1016/j.jenvp.2015.11.009

Grankvist, G., Dahlstrand, U., & Biel, A. (2004). The impact of environmental labelling on consumer preference: Negative vs. positive labels. *Journal of Consumer Policy*, *27*(2), 213–230. https://doi.org/10.1023/B:COPO.0000028167.54739.94

Heinzle, S. L., & Wüstenhagen, R. (2012). Dynamic adjustment of eco-labeling schemes and consumer choice—The revision of the EU energy label as a missed opportunity? *Business Strategy and the Environment*, *21*(1), 60–70. https://doi.org/10.1002/bse.722

Holmgren, M., Andersson, H., & Sörqvist, P. (2018). Averaging bias in environmental impact estimates: Evidence from the negative footprint illusion. *Journal of Environmental Psychology*, *55*, 48–52. https://doi.org/10.1016/j.jenvp.2017.12.005

Hunt, R. R. (2006). The concept of distinctiveness in memory research. In J. B. Worthen & R. R. Hunt (Eds.), *Distinctiveness and memory* (pp. 3–25). Oxford University Press. https://doi.org/10.1093/acprof:oso/9780195169669.003.0001

Hunt, R. R., & Smith, R. E. (1996). Accessing the particular from the general: The power of distinctiveness in the context of organization. *Memory & Cognition*, *24*(2), 217–225. https://doi.org/10.3758/BF03200882

Logie, R. H., Del Sala, S., Wynn, V., & Baddeley, A. D. (2000). Visual similarity effects in immediate verbal serial recall. *The Quarterly Journal of Experimental Psychology Section A*, *53*(3), 626–646. https://doi.org/10.1080/713755916

Lurie, N. H., & Mason, C. H. (2007). Visual representation: Implications for decision making. *Journal of Marketing*, *71*(1), 160–177. https://doi.org/10.1509/jmkg.71.1.160

Mäntylä, T., & Nilsson, L.-G. (1988). Cue distinctiveness and forgetting: Effectiveness of self-generated retrieval cues in delayed recall. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *14*(3), 502–509. https://doi.org/10.1037/0278-7393.14.3.502

Nairne, J. S. (1990). A feature model of immediate memory. *Memory & Cognition*, *18*(3), 251–269. https://doi.org/10.3758/BF03213879

Newell, R. G., & Siikamäki, J. (2014). Nudging energy efficiency behavior: The role of information labels. *Journal of the Association of Environmental and Resource Economists*, *1*(4), 555–598. https://doi.org/10.1086/679281

Oberauer, K. (2009). Interference between storage and processing in working memory: Feature overwriting, not similarity-based competition. *Memory & Cognition*, *37*(3), 346–357. https://doi.org/10.3758/MC.37.3.346

Oberauer, K., & Kliegl, R. (2006). A formal model of capacity limits in working memory. *Journal of Memory and Language*, *55*(4), 601–626. https://doi.org/10.1016/j.jml.2006.08.009

Peirce, J., Gray, J. R., Simpson, S., MacAskill, M., Höchenberger, R., Sogo, H., Kastman, E., & Lindeløv, J. K. (2019). PsychoPy2: Experiments in behavior made easy. *Behavior Research Methods*, *51*(1), 195–203. https://doi.org/10.3758/s13428-018-01193-y

Sammer, K., & Wüstenhagen, R. (2006). The influence of eco-labelling on consumer behaviour—Results of a discrete choice analysis for washing machines. *Business Strategy and the Environment*, *15*(3), 185–199. https://doi.org/10.1002/bse.522

Schmidt, S. R. (1985). Encoding and retrieval processes in the memory for conceptually distinctive events. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *11*(3), 565–578. https://doi.org/10.1037/0278-7393.11.3.565

Shiffrin, R. M., & Steyvers, M. (1997). A model for recognition memory: REM—Retrieving effectively from memory. *Psychonomic Bulletin & Review*, *4*(2), 145–166. https://doi.org/10.3758/BF03209391

Skog, E., & Sörqvist, P. (2026). Consumers' understanding of energy labels: Perception of eco-design with scale range and color-coding. *Psychology & Marketing*, *43*(4), 934–952. https://doi.org/10.1002/mar.70095

Skourtos, M., Damigos, D., Tourkolias, C., & Kontogianni, A. (2021). Efficient energy labelling: The impact of information content and style on product choice. *Energy Efficiency*, *14*(6), Article 58. https://doi.org/10.1007/s12053-021-09950-3

Smith, R. E., & Hunt, R. R. (2000). The effects of distinctiveness require reinstatement of organization: The importance of intentional memory instructions. *Journal of Memory and Language*, *43*(3), 431–446. https://doi.org/10.1006/jmla.2000.2707

Sörqvist, P., Skog, E., Heidenreich, J., & Marsh, J. E. (2025). All's eco-friendly that ends eco-friendly (if remembered as such): Memory processes in retrospective judgment of environmentally significant sequences. *Applied Cognitive Psychology*, *39*(4), Article e70103. https://doi.org/10.1002/acp.70103

Stadelmann, M., & Schubert, R. (2018). How do different designs of energy labels influence purchases of household appliances? A field study in Switzerland. *Ecological Economics*, *144*, 112–123. https://doi.org/10.1016/j.ecolecon.2017.07.031

Stasiuk, K., & Maison, D. (2022). The influence of new and old energy labels on consumer judgements and decisions about household appliances. *Energies*, *15*(4), Article 1260. https://doi.org/10.3390/en15041260

Thøgersen, J., Dessart, F. J., Marandola, G., & Hille, S. L. (2024). Positive, negative or graded sustainability labelling? Which is most effective at promoting a shift towards more sustainable product choices? *Business Strategy and the Environment*, *33*(7), 6795–6813. https://doi.org/10.1002/bse.3838

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, *185*(4157), 1124–1131. https://doi.org/10.1126/science.185.4157.1124

van den Broek, K. L. (2019). Household energy literacy: A critical review and a conceptual typology. *Energy Research & Social Science*, *57*, Article 101256. https://doi.org/10.1016/j.erss.2019.101256

Waddill, P. J., & McDaniel, M. A. (1998). Distinctiveness effects in recall: Differential processing or privileged retrieval? *Memory & Cognition*, *26*(1), 108–120. https://doi.org/10.3758/BF03211374

Wang, Z., Sun, Q., Wang, B., & Zhang, B. (2019). Purchasing intentions of Chinese consumers on energy-efficient appliances: Is the energy efficiency label effective? *Journal of Cleaner Production*, *238*, Article 117896. https://doi.org/10.1016/j.jclepro.2019.117896

Wickelgren, W. A. (1965). Acoustic similarity and intrusion errors in short-term memory. *Journal of Experimental Psychology*, *70*(1), 102–108. https://doi.org/10.1037/h0022015

Received November 20, 2025
Revision received February 6, 2026
Accepted April 6, 2026 ■