Psychonomic Bulletin & Review (2022) 29:1035–1044
https://doi.org/10.3758/s13423-021-02031-7

BRIEF REPORT

Transfer of category learning to impoverished contexts

Peter S. Whitehead[1,2] · Amanda Zamary[2] · Elizabeth J. Marsh[2]

Accepted: 14 October 2021 / Published online: 16 December 2021

Abstract
Learning often happens in ideal conditions, but then must be applied in less-than-ideal conditions – such as when a learner studies clearly illustrated examples of rocks in a book but then must identify them in a muddy field. Here we examine whether the benefits of interleaving (vs. blocking) study schedules, as well as the use of feature descriptions, supports the transfer of category learning in new, impoverished contexts. Specifically, keeping the study conditions constant, we evaluated learners' ability to classify new exemplars in the same neutral context versus in impoverished contexts in which certain stimulus features are occluded. Over two experiments, we demonstrate that performance in new, impoverished contexts during test is greater for participants who received an interleaved (vs. blocked) study schedule, both for novel and for studied exemplars. Additionally, we show that this benefit extends to both a short (3-min) or long (48-h) test delay. The presence of feature descriptions during learning had no impact on transfer. Together, these results extend the growing literature investigating how changes in context during category learning or test impacts performance and provide support for the use of interleaving to promote the *far transfer* of category knowledge to impoverished contexts.

Keywords Category learning · Transfer · Education · Interleaving · Feature descriptions

Introduction

Learning often occurs under pristine conditions, such as when a geologist-in-training studies an illustrated guide to rocks, a birder listens to labeled recordings of bird calls, or a student reviews stars and constellations in a diagram of the night sky. Clear and complete (often decontextualized) examples are the norm in guidebooks and textbooks, as well as in many psychology experiments. Robust learning, however, must transfer to different, often less than ideal conditions. Oftentimes rocks are muddy, bird calls are intermingled with other sounds, and light pollution obscures the night sky. Here we examine whether strategies known to promote category learning also support transfer of that knowledge to the identification of new exemplars in impoverished contexts.

At the broadest level, transfer refers to learning that persists despite differences at test, versus encoding, in the items – the *content* of what is learned – or conditions – the *context* of learning (see Barnett & Ceci, 2002; Taatgen, 2013). *Close transfer* involves similar items or conditions at study and test (e.g., identifying a new photo of a rock after studying guidebook photos) whereas *far transfer* involves very different items or conditions between study and test. The category learning literature has typically defined transfer as people's ability to classify new, non-studied exemplars (e.g., classifying new rock as granite), i.e., relatively *close transfer*.

A large literature has revealed many principles that promote this form of transfer. For example, people correctly classify more new exemplars following interleaved learning, where exemplars from different categories are intermixed, as opposed to blocked study, where exemplars from the same category are studied in sequence (for a recent review and meta-analysis, see Dunlosky et al., 2013; see also Brunmair & Richter, 2019; Kornell & Bjork, 2008; cf. Flesch et al., 2018; Goldstone, 1996; Tauber et al., 2013). Furthermore, transfer is improved following distributed practice – the spreading of study opportunities over a greater period of

Peter S. Whitehead and Amanda Zamary are co-first authors

✉ Peter S. Whitehead
peter.whitehead@duke.edu

[1] Center for Cognitive Neuroscience, Duke University, LSRC, Box 90999, Durham, NC 27708, USA

[2] Department of Psychology & Neuroscience, Duke University, Durham, NC 27708, USA

Fig. 1 An example (**a**) of interleaved and blocked study sequence. The above example uses a reduced number of exemplars than actually used in Experiments 1 and 2 for visualization purposes. In (**b**) an example of the same rock exemplar in both a control and impoverished context. In (**c**) an example of the rocks when studied under the inclusion of feature descriptions condition

time – than temporally massed practice (for review, see Benjamin & Tullis, 2010; Cepeda et al., 2006; Dunlosky et al., 2013). Transfer also improves following study of exemplars labeled with items' key diagnostic features, as compared to studying unlabeled exemplars (Miyatsu et al., 2019). Many of the principles that support transfer of categorical knowledge are the same as ones that support memory. For example, learning through attempted labeling (with feedback) is more effective than studying labeled exemplars (Levering & Kurtz, 2015), paralleling the large literature showing that retrieval practice is a more effective learning strategy than reading (Roediger & Butler, 2011).

Transfer of category learning in the real world, however, is rarely as forgiving as in a textbook – perceptual and cognitive demands increase in cases of *far transfer*, where test exemplars and contexts differ greatly from learning. For example, naturally encountered pieces of granite differ in shape and size, can be broken or partially obscured, and appear different depending on lighting or environmental factors (mud, dirt, etc.). These perceptual obfuscations of stimuli features can indiscriminately affect both discriminative and characteristic features of stimuli, altering between-category differences as well as within-category similarities (Carvalho & Goldstone, 2017). To date, a small, but growing, literature has examined how people learn to categorize when *learning* is less than ideal, examining the occlusion of study exemplars' perceptual features (e.g., Hornsby & Love, 2014; Meagher et al., 2018; Taylor & Ross, 2009) or restricting the training range to typical cases (Hornsby & Love, 2014). But no studies have examined difficulties introduced at test, which may or may not have similar effects as those observed during learning – the memory, for example, is more affected by divided attention at study than at test (Craik et al., 1996).

Here we focus on the potential benefits of interleaved study, as opposed to blocked study (see Dunlosky et al., 2013). To examine transfer in impoverished contexts, we simulated a real-world impoverished context common in aviation and maritime operations: night vision (Gauthier et al., 2008; Johnson, 2004; Ruffner et al., 2001; Salazar et al., 2003). We adapted rock stimuli (Miyatsu et al., 2019) to simulate their appearance via night goggles, partially occluding two diagnostic features: color and, to a lesser extent, granularity (see Fig. 1). Thus, this night goggle simulation affected both *discriminative features* (ones that differentiate between categories) and *characteristic features* (those shared within a category). For example, color is discriminative when identifying rocky gypsum (almost always white) and obsidian (almost always black), as those colors are relatively unique in the set of rocks used. In contrast, sandstone rocks are similarly colored to each other, but their color palette is also similar to that of many other rocks – making color characteristic but not discriminative (see Carvalho & Goldstone, 2017; Nosofsky et al., 2017).

Compared to blocked study, interleaving often leads to better learning (Dunlosky et al., 2013). This is especially true for highly overlapping categories (Carvalho & Goldstone, 2014) and more-to-difficult-to-learn categories (Zulkiply & Burt, 2013). Sequential Attention Theory (SAT) posits that such effects occur because interleaving directs attention to features that differentiate between categories (*discriminative* features), whereas blocked learning highlights shared features within a category (*characteristic* features; Carvalho & Goldstone, 2015; Carvalho & Goldstone, 2017). While task specific circumstances and variability influence the effectiveness of each study strategy, one prediction of SAT, however, is that a *blocked* study schedule may be a more effective learning strategy when the nature of the final test is unknown (Carvalho & Goldstone, 2015). This prediction arises as SAT posits that *blocked* sequences lead to more localized representations of categories absent inter-category context. Conversely, an *interleaved* sequence contextualizes category representations, promoting interconnected, between-category representations through the learning of discriminative features. Simply put, the larger learning context exerts less influence on representations extracted during *blocked* learning than *interleaved* learning, predicting that leading blocked learning will yield representations that are more context flexible and more useful when test conditions are unknown during learning. Therefore, one prediction of prediction of SAT is that *blocked* learning will benefit far transfer to a testing environment that differs from the context during learning – here, where the perceptual obfuscations indiscriminately affect both discriminative and characteristic features (Carvalho & Goldstone, 2015).

Therefore, in two studies where learning occurred under ideal, unaltered conditions, we manipulated two strategies proposed to improve learning and transfer of rock classification: interleaving (vs. blocked study) and, in an exploratory manipulation, feature descriptions, which described and circled rocks' key features (Miyatsu et al., 2019). Here we aimed to replicate two findings (pre-registered): that under standard test conditions – i.e., *control contexts* – both memory for studied items and identification of novel rocks would be greater following interleaved vs. blocked practice (see Dunlosky et al., 2013). Further, we pre-registered two novel questions: Under *impoverished contexts*, will (1) memory for studied instances and (2) transfer to new exemplars be greater following interleaved versus blocked practice? As an exploratory question (pre-registered), we manipulated whether features were labeled during learning in order to investigate whether any of the above questions are modulated by the use of feature descriptions during learning (see Miyatsu et al., 2019).

We investigated these questions in two experiments which differed only in the amount of time separating study and test. The test occurred almost immediately after study in Experiment 1, but was delayed 2 days in Experiment 2 as research on the relationship between study-test delay and the benefits of interleaving has produced mixed findings (for reviews, see Brunmair & Richter, 2019; Dunlosky et al., 2013).

Experiment 1

Method

Participants and design Duke University's internal review board approved both experiments. Both were preregistered on the Open Science Framework (https://osf.io/3fmxj/[1]). We conducted an a priori power analysis using G*Power 3.1.9.2. (Faul et al., 2007) for 2 (between: interleaved vs. blocked) × 2 (between: feature descriptions vs. no feature descriptions) ANOVAs with power set at .9, $\alpha = .05$, and Cohen's $f = .20$, which suggested a sample size of 265. Our targeted sample size was 280 participants, given that we expected some level of attrition and non-compliance.

In total, Experiment 1 included 280 participants recruited from Prolific in February/March 2020 (www.prolific.co; Palan & Schitter, 2018). Participants were required to (1) currently reside in the USA or UK, (2) speak English as a first language, (3) have a 90% study approval rate, (4) have a minimum of 100 study submissions, and (5) complete the study using a desktop computer. Data for 31 participants were excluded from analysis due to failure to complete the full study ($n = 5$) or failure of one or both attention checks ($n = 26$). Thus, the final sample included 249 participants (M age = 38 years, SD = 13; 57% female; 87% White).

Materials Materials included images of rocks from Miyatsu et al. (2019) from 12 different rock categories (i.e., amphibolite, breccia, conglomerate, gneiss, granite, obsidian, marble, pegmatite, pumice, rock gypsum, sandstone, and slate). Out of the 144 exemplars used in Miyatsu et al. (2019), we randomly selected 120 exemplars for the current research (ten exemplars per category). Across all participants, 72 of these exemplars were randomly selected to serve as study stimuli and the remaining 48 served as novel rocks on the final classification test.

At study, images were presented either with or without highlighted feature descriptions, depending on group

[1] We also ran Experiment 1 on Amazon's MTurk. However, 85 out of the 280 participants (30%) of the sample either (a) submitted poor-quality data or (b) were identified as bots. In hindsight, we realized that our exclusion criteria may have not been strong enough given the increase in bots on Mturk and we are not confident in the reliability of this dataset. Therefore, we do not present these data here.

assignment. Images with feature descriptions included the rock with key category features circled and notated, whereas images without feature descriptions did not (see Miyatsu et al., 2019, Exp. 2). At test, all images were presented without feature descriptions. Out of the 48 studied rocks, 24 were randomly selected to be tested in the original control context and another unique set of 24 were tested in the impoverished context. Similarly, out of the 48 novel rocks (the transfer items), we randomly selected 24 to be tested in the original neutral context and 24 to be tested in the impoverished context. Rocks tested in the neutral context were presented in their original condition in full color (as in Miyatsu et al., 2019). For impoverished context test stimuli, we adapted images from Miyatsu et al. (2019) to appear as if they were being viewed under a night vision filter. This was done using the following procedure. First, each image was converted to a monochrome scale. Then the mixing of RGB source channels was altered to be 0%, 100%, and 0%, respectively, for each channel. Images where then altered along HSL color lines to have the values 75, 100, and -25, respectively, for each color property in HSL space. Finally, for each image, the black value in CMYK color space was altered to be 100%. This effectively simulated each rock under a night vision context. The use of a "night vision" filter therefore does not simply darken stimuli, but mimics the complex and multidimensional way lighting and other environmental features might influence the color of real-world stimuli. For these adapted stimuli, please see https://osf.io/3fmxj/.

Procedure After reserving a spot in the study on Prolific, participants were redirected to Gorilla (www.gorilla.sc), an online experiment builder that presented all tasks and instructions (Anwyl-Irvine et al., 2020). Participants then completed the consent form and were asked to minimize distractions prior to beginning the experiment. Next, participants were told that they would be asked to learn 12 types of rocks and that the experiment would have two major phases (i.e., a learning phase and a testing phase).

During the learning phase, participants studied 6 different exemplars of each rock category; they saw each example twice during the study phase, equating to 144 study trials in total. On each study phase trial, participants passively viewed a rock image presented with its category name for 6 s (either with or without feature descriptions, depending on group assignment). Images were presented in either a blocked or interleaved study schedule, depending on group assignment. For each group, we created a fixed, randomized order for study presentation (see Fig. 1). In the blocked group, participants studied all six exemplars for a given category back-to-back in the same order twice prior to moving onto the next category. In the interleaved group, participants studied one exemplar per category in six blocks of 12. Once all six blocks were presented, participants received the six blocks for a second round of study. At least three exemplars from different categories were presented prior to receiving a new instance from the same category.

In both groups, two pictures of unrelated objects (i.e., a stack of books and a fork) served as attention check trials. These trials were presented at the ends of blocks 4 and 11 of study (after trials 48 and 132; see Meade & Craig, 2012, for recommendations to use one attention check item per every 50–100 trials in survey research). On each attention check trial, the image and its name were also presented for 6 s. Participants were then immediately prompted to type in the name of the object on the next screen prior to continuing the study. Participants were unaware that these attention check trials would occur. After completing the learning phase, participants played *Tetris* for 3 min as a distractor task.

During the testing phase, participants took classification tests that required them to classify each of a series of rocks by selecting from a list of 12 possibilities (presented alphabetically). Participants indicated their choice by clicking on it with the mouse cursor. Participants were asked to avoid outside sources and to simply try their best. Participants were not instructed that some rocks would be new and some would be those they previously studied. The test was self-paced.

The standard and impoverished context tests were counterbalanced evenly within each group. For both tests, novel items were always tested prior to studied items, but each half of the tests was randomized anew for each participant. For the impoverished context test, participants were informed that each rock would be presented under a night vision filter during classification. After completing the final tests, participants filled out a demographics questionnaire and were awarded $6.75 for their participation.

Analysis All primary analyses for both experiments were pre-registered at https://osf.io/3fmxj/. Our pre-registered analysis plan stated that we would conduct four 2 × 2 between-subjects ANOVAs (study schedule × feature descriptions) for each of the example classification item types (i.e., studied items in same context, studied items in different context, novel items in same context, novel items in different context). However, upon inspection of the data, we deviated from our pre-registered analysis plan in favor of a linear mixed-models approach.

Classification accuracy data from the testing phase were fit to five generalized mixed effects logit models using the *lme4* packages in R (Bates et al., 2015). Each model had an identical crossed random-effects structure, with a random intercept for each subject, as well as each rock category type. The first, null model included only the random-effects structure.

Table 1 Means and standard errors for final test performance for Experiment 1 (3-min delay)

Study characteristics		Test characteristics			
Study schedule	Feature descriptions	Control context		Impoverished context	
		Studied	Novel	Studied	Novel
Interleaved	Yes	68 (2)	59 (2)	57 (2)	52 (2)
Interleaved	No	68 (2)	57 (2)	56 (2)	51 (2)
Blocked	Yes	56 (2)	50 (2)	45 (2)	44 (2)
Blocked	No	60 (2)	51 (2)	48 (2)	48 (2)

Note: *N*s for each of the groups were as follows: Interleaved Feature Descriptions (60), Interleaved No Feature Descriptions (59), Blocked Feature Descriptions (65), Blocked No Feature Descriptions (65). Numbers listed are percentages. Standard errors are presented in parentheses

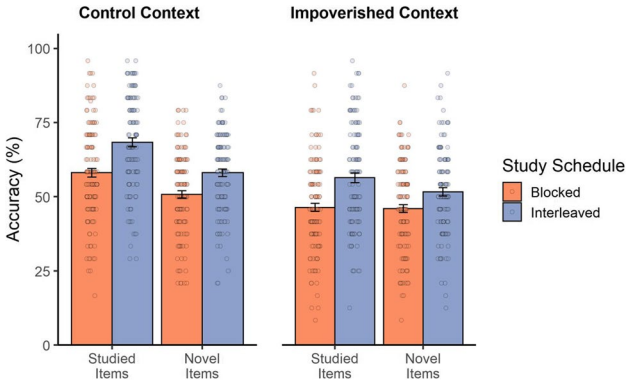

Fig. 2 Categorization accuracy (%) in Experiment 1, as a function of the between-participants study schedule, context (control vs. impoverished), and whether items were previously studied or novel. Error bars are standard error, and data points are individual participants

Our hypotheses focused on the effects of four fixed factors on categorization performance at test: Study Schedule (Interleaved vs. Blocked study schedule), Study Status of test items (Studied vs. Novel exemplars), Context (standard vs. Impoverished contexts), and Feature Descriptions (Present vs. Absent at study). We structured a set of four hierarchical generalized mixed effects models to determine the validity of including each of these fixed effects. We iteratively added each fixed factor to successive models in order to test whether their inclusion improved model fit, and indicating whether that factor provided significant explanatory power in characterizing categorization performance at test. The fixed-effects structure of these models can be summarized as: Null: Random effects only, Model 1: Study Schedule, Model 2: Study Schedule ×Study Status, Model 3: Study Schedule ×Study Status × Context, Model 4: Study Schedule ×Study Status × Context + Feature Description. The fit of these mixed models was determined using the *anova()* command in R to calculate AIC scores and conduct a chi-squared test of each model against its hierarchically subordinate model (i.e., null vs. 1-factor model).

Results

Means and standard errors for all groups are presented in Table 1. These results are visualized in Fig. 2. After data were fit to each model, the model fit test indicated that Model 3, in which the Study Schedule, Study Status, and Context factors were included as main effects and interactions was the best fitting model ($\Delta AIC = 276.28$; $p < .0001$; Table 2). The inclusion of the Feature Description factor in Model 4 as a main effect did not significantly improve model fit ($\Delta AIC = -1.79$; $p = .643$; Table 2). As such, we concluded that the inclusion (or lack thereof) of Feature Descriptions during study did not impact subsequent memory or transfer performance, and no follow-up tests of interactions were considered. The results of Model 3 can be seen in Table 3.

To explicitly test our four pre-registered hypotheses, we performed follow-up contrasts of Model 3. We found that interleaved versus blocked study schedules led to greater classification accuracy in *control contexts* for studied items, as well as novel transfer items (Memory: $\beta = -0.60$, $p < .0001$; Transfer: $\beta = -0.41$, $p = .0002$). This confirmed our first two hypotheses, and replicated previous findings (see Dunlosky et al., 2013). Our latter two hypotheses focused on the benefits of interleaving for transfer and memory in *impoverished contexts.* Here, follow-up contrasts of Model 3 demonstrated that interleaved versus blocked study schedules led to greater classification accuracy in *impoverished contexts* for studied, memory items, as well as novel, transfer items (Memory: $\beta = -0.55$, $p < .0001$; Transfer: $\beta = -0.31$, $p = .0051$). These results demonstrate for the first time that an interleaved study schedule benefitted later categorization in an *impoverished context,* for both studied and novel (transfer) rocks.

Experiment 2

Experiment 2 was designed to replicate findings from Experiment 1 with one major change – the time between study and test was extended from 3 min to 48 h. Some effects in cognitive science depend upon the length of the delay between learning and testing. For instance, the benefits of retrieval practice (see Congleton & Rajaram, 2012; Karpicke & Roediger, 2007; Roediger & Karpicke, 2006) are normally observed on delayed tests (e.g., 2 days), whereas the effect disappears or even reverses in favor of rereading on relatively immediate tests (e.g., 3 min).Existing research

Table 2 Results of the model comparison for hierarchical models of classification accuracy in Experiment 1

	Parameters	AIC	logLik	Chi-squared	*df*	*p*
Null	3	26325	-13160			
×Study Schedule	4	26304.95	-13148	22.05	1	**<.0001**
×Study Status	6	26192.92	-13090	116.03	2	**<.0001**
× Context	10	25916.64	-12948	284.28	4	**<.0001**
+ Feature Description	11	25918.43	-12948	0.21	1	0.643

df degrees of freedom

Table 3 Summary results of the Study Schedule × Study Status × Context model for Experiment 1

	OR	SE	*z*	*p*
Intercept	0.511	0.353	1.45	0.148
Study Schedule	0.597	0.111	5.38	**<0.001**
Study Status	-0.403	0.060	-6.77	**<0.001**
Context	-0.642	0.060	-10.72	**<0.001**
Study Schedule × Study Status	-0.190	0.087	-2.18	**0.029**
Study Schedule × Context	-0.045	0.087	-0.52	0.603
Study Status × Context	0.379	0.085	4.49	**<0.001**
Study Schedule × Study Status × Context	-0.052	0.123	-0.42	0.672

OR odds ratio, *SE* standard error

Bolded values indicate significance

Table 4 Means and standard errors for final test performance for Experiment 2 (2-day delay)

Study characteristics		Test characteristics			
Study schedule	Feature descriptions	Control context		Impoverished context	
		Studied	Novel	Studied	Novel
Interleaved	Yes	66 (2)	55 (2)	54 (2)	52 (2)
Interleaved	No	64 (3)	54 (2)	53 (2)	50 (2)
Blocked	Yes	54 (2)	46 (2)	42 (2)	46 (2)
Blocked	No	54 (3)	44 (2)	45 (3)	44 (2)

Note: *N*s for each of the groups were as follows: Interleaved Feature Descriptions (61), Interleaved No Feature Descriptions (58), Blocked Feature Descriptions (60), Blocked No Feature Descriptions (61). Numbers listed are percentages. Standard errors are presented in parentheses

evaluating the extent to which interleaving benefits depend on test delay is mixed (for reviews, see Brunmair & Richter, 2019; Dunlosky et al., 2013). A recent meta-analysis by Brunmair and Richter (2019) suggests that interleaving benefits are not moderated by test delay or whether test items are studied or novel items. However, given that transfer across different contexts is vastly understudied, the extent to which this meta-analysis applies to the current research is unclear. Thus, we increased the delay between study and test to be approximately 48 h in Experiment 2 to evaluate the extent to which effects and/or effect sizes may change at longer delays.

Method

Participants and design As in Experiment 1, our targeted sample size was 265 participants. However, we again oversampled and recruited 276 participants from Prolific in May/June 2020, given that we expected some level of attrition and non-compliance. Eligibility requirements were the same as Experiment 1. Data for 30 participants were excluded from analysis due to failure to complete the full study ($n = 4$) or failure of one or both attention checks ($n = 32$). Thus, the final sample included 240 participants (M age = 36 years, SD = 13; 69% female; 88% White).

Procedure The procedure for Experiment 2 was exactly the same as Experiment 1, except for one change. After participants completed study, they did not play Tetris. Instead, they were told that they were done for the day and would receive a reminder email in approximately 48 h to complete the second part of the study (i.e., the testing phase). Upon completion of both sessions, participants were awarded $10.00 for their participation.

Analysis As in Experiment 1, we deviated from our preregistered analysis plan and fit the classification accuracy data during the testing phase to a set of generalized linear mixed logit models.

Results

Means and standard errors for all groups are presented in Table 4. These results are visualized in Fig. 3. After data were fit to each model, the model fit test indicated that Model 3, in which the Study Schedule, Study Status, and Context factors were included as main effects and interactions was the best fitting model (ΔAIC = 199; $p <$ .001; Table 5). The inclusion of the Feature Description factor in Model 4 as a main effect did not significantly

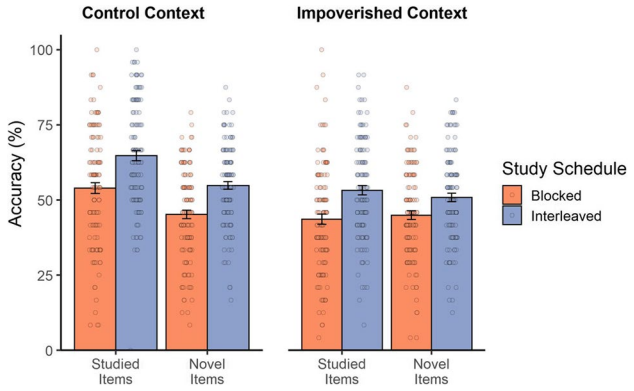

Fig. 3 Categorization accuracy (%) in Experiment 2, as a function of the between-participants study schedule, context (control vs. impoverished), and whether items were previously studied or novel. Error bars are standard error, and data points are individual participants

We again explicitly tested our four hypotheses by conducting follow-up contrasts of Model 3. Interleaved study again led to better classification performance than did blocked study, regardless of whether the test items were studied ones or novel, transfer items (Memory: $\beta = -0.62$, $p < .0001$; Transfer: $\beta = -0.55$, $p < .0001$). Further, interleaved vs. blocked study schedules led to greater classification accuracy in *impoverished contexts,* for both studied items as well as novel, transfer items (Memory: $\beta = -0.56$, $p < .0001$; Transfer: $\beta = -0.35$, $p = .0046$). Importantly, these results replicate and extend our findings from Experiment 1, indicating that the benefits of interleaved study lead to more accurate classification of studied and novel rocks tested in *impoverished contexts,* up to ~48 h after study.

Table 5 Results of the model comparison for hierarchical models of classification accuracy in Experiment 2

	Parameters	AIC	logLik	Chi-squared	*df*	*p*
Null	3	25292	-12643			
×Study Schedule	4	25272	-12632	21.45	1	**<.0001**
×Study Status	6	25194	-12591	81.94	2	**<.0001**
× Context	10	24995	-12487	207.59	4	**<.0001**
+ Feature Description	11	24996	-12487	0.22	1	0.639

df degrees of freedom

Bolded values indicated significance

Table 6 Summary results of the Study Schedule × Study Status × Context model for Experiment 2

	OR	SE	*z*	*p*
Intercept	0.261	0.343	0.76	0.447
Study Schedule	0.615	0.122	5.04	**<.001**
Study Status	-0.497	0.063	-7.95	**<.001**
Context	-0.590	0.063	-9.41	**<.001**
Study Schedule × Study Status	-0.064	0.089	-0.72	0.474
Study Schedule × Context	-0.060	0.089	-0.68	0.500
Study Status × Context	0.576	0.089	6.50	**<.001**
Study Schedule × Study Status × Context	-0.146	0.125	-1.17	0.243

OR odds ratio, *SE* standard error

Bolded values indicate significance

improve model fit (ΔAIC = -1; $p = .639$; Table 5). As such, we concluded that the inclusion (or lack thereof) of Feature Descriptions during study did not impact subsequent memory or transfer performance, and no follow-up tests of interactions were considered. The results of Model 3 can be seen in Table 6.

General discussion

Across two experiments, we demonstrated far transfer: participants successfully identified novel rocks in a simulated night vision environment that obscured rock color and granularity. Performance (Exp. 1: 50%; Exp. 2: 48%) was much higher than chance (8.33%). Critically, interleaving (vs. blocked study) led to better identification of novel, transfer items in this impoverished context. Interleaving also benefited the identification of studied rocks in the impoverished context (Figs. 2 and 3, Tables 1 and 4). As expected, we replicated previous work showing the benefits of interleaving in a typical test environment, where no features were obscured. All benefits of interleaving occurred both immediately and after a 2-day delay, congruent with a recent meta-analysis from Brunmair and Richter (2019) that suggested the benefits from an interleaved study schedule are not dependent on the length of test delay. Unrelated to our pre-registered hypotheses, we also saw a consistent significant interaction between old/new (study) status and context. Typically, people are much better at classifying studied items than novel transfer ones – a finding we observe under control contexts

at test. But, this benefit was reduced under impoverished contexts at test too; performance on studied and novel items was similar (Figs. 2 and 3, Tables 3 and 6).

In contrast, feature descriptions (or lack thereof) did not affect later transfer (Tables 1, 2, 3, 4, 5 and 6), a finding inconsistent with Miyatsu et al. (2019, Experiments 2 and 3). There, labeling features during learning improved performance on a standard transfer test 2 days later (where required participants to identify new exemplars, but in the same unobstructed context as studied items). In our work, feature descriptions had almost no effect on performance regardless of final test context or item type (studied vs. novel). To be very clear, this is not to say the present work finds evidence *against* the benefits of feature descriptions during study. Feature highlighting is thought to benefit category learning through the biasing of attention toward category-relevant features or via the promotion of learning qualitative difference in category representations (Miyatsu et al., 2019). The inclusion here of other study strategies (i.e., interleaving vs. blocked study schedules) is not necessarily additive with the inclusion (or not) of feature descriptions. Therefore, the benefits of feature descriptions during study may be rendered ineffective when paired with other study strategies.

To our knowledge, this is the first demonstration that interleaving (vs. blocked) study schedules lead to better categorization of novel exemplars in a new, impoverished contexts. This finding adds to a growing literature focused on how changes in perceptually available features between learning and test impacts performance (Hornsby & Love, 2014; Meagher et al., 2018; Taylor & Ross, 2009). While previous literature has investigated the effects of manipulating available perceptual features during *learning* (see Meagher et al., 2018), understanding how changes in the testing environment affect transfer (i.e., where perceptual features of stimuli are occluded) remains understudied (see Hornsby & Love, 2014).

These results are inconsistent with the prediction of the SAT on the ideal study strategy when test conditions are unknown. The SAT suggests that blocked learning is more advantageous under transfer conditions at test that differ from study conditions. The SAT posits that different learning strategies draw attention to different features of to-be-learned exemplars, with consequences for later performance. Interleaving promotes the learning of differences between exemplars coming from different categories, while blocking promotes the learning of similarities between items of the same category (Dunlosky et al., 2013; Carvalho & Goldstone, 2017; Nosofsky, 2011). Given the benefits of interleaving observed in both of our experiments, the learning of discriminative features appears to matter more when the goal is transfer in a situation in which discriminative and characteristic features are both occluded (see Carvalho & Goldstone, 2017, Figs. 2 and 3). This result is inconsistent with this prediction of SAT; instead, the creation of robust, inter-related networks of categorization via learning discriminative features during *interleaved* study may be more resilient to broad contextual changes, whereas locally segregated networks of characteristic within-category features promoted by blocked study might be less adaptable (Zulkiply & Burt, 2013; see also Goldstone, 1996).

The current results could alternatively be explained via differences in the initial, baseline quality of learning. While we do not have a direct measure of initial learning (as subjects studied image-label pairs without making any responses), it is reasonable to assume learning was higher in the interleaved condition given past research. In both studies (Tables 3 and 6) interleaved study always led to better categorization performance, for both studied and novel items. However, the data observed in the impoverished condition do not perfectly mirror those in the control context (as might be expected if the amount of learning at the end of the learning phase was the main predictor of final test performance). Interleaving led to better performance in the impoverished condition, with the expected decline for new exemplars (as compared to studied rocks; Figs. 2 and 3). Importantly, this was not true for a *blocked* study schedule; performance between *studied* and *novel* items was quite similar, albeit low, in the *impoverished* context (Figs. 2 and 3). Perhaps a blocked study schedule can protect against a decline in categorization performance for novel items, in comparison to studied items; however, content is simply not learned as well overall as in an *interleaved* study context. Conversely, this difference could also be interpreted as a memory boost specifically for studied items learned under an interleaved schedule, as a result of the benefit of distributed practice on memory retention (for review, see Dunlosky et al., 2013). However, whether these effects are present under less passive study conditions remains to be studied (see Carvalho & Goldstone, 2015).

While previous literature has explored the *far transfer* of category learning for relational category structures (Patterson & Kurtz, 2020; see also Lowenstein, 2010), these studies often conceptualize *far transfer* as simply obscuring only characteristic features of stimuli, while leaving intact the diagnostic and relational features. Here, our conceptualization of the contextual *far transfer* of category learning seeks to account for the perceptual conditions of new, real-world contexts which often indiscriminately obscure both characteristic and diagnostic features of an item. A salient, but imperfect, analogy might be how the change in light during sunset can alter the visual color of the entire rockface of a mountain range, as opposed to changing a specific set of features – i.e., altering all colors of a granite rockface instead of only the red specks in the granite. For example, the rockfaces of the Sandia Mountain range outside Albuquerque, New

Mexico change color in the evening light without regard to the construct of discriminative or characteristic features. Together, we propose that these results speak both to current debates within the literature regarding the usefulness of blocked versus interleaved learning for complex real-world stimuli (for review, see Hughes & Thomas, 2021; see also Flesch et al., 2018), as well as more generally on the usefulness of discriminative versus characteristic features for the *far transfer* of category learning in impoverished, real-world contexts (see also Murphy, 1982; Murphy & Ross, 2005).

Conclusion

The current research demonstrated the benefits of interleaving study schedules during *learning* on performance at *test* when the physical context differs from the learning context, both for studied and novel exemplars. Here, specifically, we replicated and extended the previous finding (pre-registered) that under *control context* conditions, memory and transfer for studied instances is greater following interleaved vs. blocked practice (see Dunlosky et al., 2013), showing that this benefit also extended to new, *impoverished contexts.* While we also manipulated the use of feature descriptions during study, in order to investigate potential performance benefits at test, we found no effect of feature descriptions (or lack thereof) on performance at any level. The current work highlights the importance of manipulations at *learning* for the promotion of transfer of category learning at *test* – showing that interleaving study schedules promotes better *far transfer* of category learning. Future work should continue to pursue investigations of the benefits of different learning strategies and manipulations to promote the *far transfer* of category learning.

Acknowledgements This work was supported by ONR N00014-18-1-2871.

Data availability Data and materials can be accessed at: https://osf.io/3fmxj/

References

Anwyl-Irvine, A. L., Massonnié, J., Flitton, A., Kirkham, N., & Evershed, J. K. (2020). Gorilla in our midst: An online behavioral experiment builder. *Behavior Research Methods*, *52*(1), 388–407. https://doi.org/10.3758/s13428-019-01237-x

Barnett, S. M., & Ceci, S. J. (2002). When and where do we apply what we learn?: A taxonomy for far transfer. *Psychological Bulletin*, *128*(4), 612–637. https://doi.org/10.1037/0033-2909.128.4.612

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using **lme4**. *Journal of Statistical Software*, *67*(1). 10.18637/jss.v067.i01

Benjamin, A. S., & Tullis, J. (2010). What makes distributed practice effective? *Cognitive Psychology*, *61*(3), 228–247. https://doi.org/10.1016/j.cogpsych.2010.05.004

Brunmair, M., & Richter, T. (2019). Similarity matters: A meta-analysis of interleaved learning and its moderators. *Psychological Bulletin*, *145*(11), 1029–1052. https://doi.org/10.1037/bul0000209

Carvalho, P. F., & Goldstone, R. L. (2014). Putting category learning in order: Category structure and temporal arrangement affect the benefit of interleaved over blocked study. *Memory & Cognition*, *42*(3), 481–495. https://doi.org/10.3758/s13421-013-0371-0

Carvalho, P. F., & Goldstone, R. L. (2015). What you learn is more than what you see: What can sequencing effects tell us about inductive category learning? *Frontiers in Psychology*, *6*, 505. https://doi.org/10.3389/fpsyg.2015.00505

Carvalho, P. F., & Goldstone, R. L. (2017). The sequence of study changes what information is attended to, encoded, and remembered during category learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *43*(11), 1699–1719. https://doi.org/10.1037/xlm0000406

Cepeda, N. J., Pashler, H., Vul, E., Wixted, J. T., & Rohrer, D. (2006). Distributed practice in verbal recall tasks: A review and quantitative synthesis. *Psychological Bulletin*, *132*(3), 354–380. https://doi.org/10.1037/0033-2909.132.3.354

Congleton, A., & Rajaram, S. (2012). The origin of the interaction between learning method and delay in the testing effect: The roles of processing and conceptual retrieval organization. *Memory & Cognition*, *40*(4), 528–539. https://doi.org/10.3758/s13421-011-0168-y

Craik, F. I. M., Govoni, R., Naveh-Benjamin, M., & Anderson, N. D. (1996). The effects of divided attention on encoding and retrieval processes in human memory. *Journal of Experimental Psychology: General*, *125*(2), 159–180. https://doi.org/10.1037/0096-3445.125.2.159

Dunlosky, J., Rawson, K. A., Marsh, E. J., Nathan, M. J., & Willingham, D. T. (2013). Improving Students' Learning With Effective Learning Techniques: Promising Directions From Cognitive and Educational Psychology. *Psychological Science in the Public Interest*, *14*(1), 4–58. https://doi.org/10.1177/1529100612453266

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

Flesch, T., Balaguer, J., Dekker, R., Nili, H., & Summerfield, C. (2018). Comparing continual task learning in minds and machines. *Proceedings of the National Academy of Sciences*, *115*(44), E10313–E10322. https://doi.org/10.1073/pnas.1800755115

Gauthier, M. S., Parush, A., Macuda, T., Tang, D., Craig, G., & Jennings, S. (2008). The impact of night vision goggles on way-finding performance and the acquisition of spatial knowledge. *Human Factors, 50*(2), 311–321

Goldstone, R. L. (1996). Isolated and interrelated concepts. *Memory & Cognition*, *24*(5), 608–628. https://doi.org/10.3758/BF03201087

Hornsby, A. N., & Love, B. C. (2014). Improved classification of mammograms following idealized training. *Journal of Applied Research in Memory and Cognition*, *3*(2), 72–76. https://doi.org/10.1016/j.jarmac.2014.04.009

Hughes, G. I., & Thomas, A. K. (2021). Visual category learning: Navigating the intersection of rules and similarity. *Psychonomic Bulletin & Review*. https://doi.org/10.3758/s13423-020-01838-0

Johnson, C. W. (2004). The role of night vision equipment in military incidents and accidents. In *Human error, safety and systems development* (pp. 1-16). Springer, Boston, MA.

Karpicke, J. D., & Roediger, H. L. (2007). Repeated retrieval during learning is the key to long-term retention. *Journal of Memory and Language*, *57*(2), 151–162. https://doi.org/10.1016/j.jml.2006.09.004

Kornell, N., & Bjork, R. A. (2008). Learning Concepts and Categories: Is Spacing the "Enemy of Induction"? *Psychological Science*, *19*(6), 585–592. https://doi.org/10.1111/j.1467-9280.2008.02127.x

Levering, K. R., & Kurtz, K. J. (2015). Observation versus classification in supervised category learning. *Memory & Cognition, 43*(2), 266–282.

Loewenstein, J. (2010). How one's hook is baited matters for catching an analogy. *Psychology of Learning and Motivation, 53*, 149–182. https://doi.org/10.1016/S0079-7421(10)53004-4

Meade, A. W., & Craig, S. B. (2012). Identifying careless responses in survey data. *Psychological Methods*, *17*(3), 437–455. https://doi.org/10.1037/a0028085

Meagher, B. J., Cataldo, K., Douglas, B. J., McDaniel, M. A., & Nosofsky, R. M. (2018). Training of rock classifications: The use of computer images versus physical rock samples. *Journal of Geoscience Education*, *66*(3), 221–230. https://doi.org/10.1080/10899995.2018.1465756

Miyatsu, T., Gouravajhala, R., Nosofsky, R. M., & McDaniel, M. A. (2019). Feature highlighting enhances learning of a complex natural-science category. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *45*(1), 1–16. https://doi.org/10.1037/xlm0000538

Murphy, G. L. (1982). Cue validity and levels of categorization. *Psychological Bulletin, 91*(1), 174.

Murphy, G. L., & Ross, B. H. (2005). The two faces of typicality in category-based induction. *Cognition, 95*(2), 175–200. https://doi.org/10.1016/j.cognition.2004.01.009

Nosofsky, R. M. (2011). The generalized context model: An exemplar model of classification. *Formal Approaches in Categorization,* 18–39.

Nosofsky, R. M., Sanders, C. A., Gerdom, A., Douglas, B. J., & McDaniel, M. A. (2017). On Learning Natural-Science Categories That Violate the Family-Resemblance Principle. *Psychological Science*, *28*(1), 104–114. https://doi.org/10.1177/0956797616675636

Palan, S., & Schitter, C. (2018). Prolific.ac—A subject pool for online experiments. *Journal of Behavioral and Experimental Finance*, *17*, 22–27. https://doi.org/10.1016/j.jbef.2017.12.004

Patterson, J. D., & Kurtz, K. J. (2020). Comparison-based learning of relational categories (you'll never guess). *Journal of Experimental Psychology: Learning, Memory, and Cognition, 46*(5), 851.

Roediger, H. L., & Butler, A. C. (2011). The critical role of retrieval practice in long-term retention. *Trends in Cognitive Sciences*, *15*(1), 20–27. https://doi.org/10.1016/j.tics.2010.09.003

Roediger, H. L., & Karpicke, J. D. (2006). Test-Enhanced Learning: Taking Memory Tests Improves Long-Term Retention. *Psychological Science*, *17*(3), 249–255. https://doi.org/10.1111/j.1467-9280.2006.01693.x

Ruffner, J. W., McAnulty, D. M., Weeter, R. D., & Wightman, D. C. (2001). *Fort Rucker Field Unit* 1988-1993. Army research inst for the behavioral and social sciences Alexandria VA.

Salazar, G., Temme, L., & Antonio, J. C. (2003). Civilian use of night vision goggles. *Aviation, Space, and Environmental Medicine, 74*(1), 79–84.

Taatgen, N. A. (2013). The nature and transfer of cognitive skills. *Psychological Review*, *120*(3), 439–471. https://doi.org/10.1037/a0033138

Tauber, S. K., Dunlosky, J., Rawson, K. A., Wahlheim, C. N., & Jacoby, L. L. (2013). Self-regulated learning of a natural category: Do people interleave or block exemplars during study? *Psychonomic Bulletin & Review*, *20*(2), 356–363. https://doi.org/10.3758/s13423-012-0319-6

Taylor, E. G., & Ross, B. H. (2009). Classifying partial exemplars: Seeing less and learning more. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *35*(5), 1374–1380. https://doi.org/10.1037/a0016568

Zulkiply, N., & Burt, J. S. (2013). The exemplar interleaving effect in inductive learning: Moderation by the difficulty of category discriminations. *Memory & Cognition*, *41*(1), 16–27. https://doi.org/10.3758/s13421-012-0238-9