Journal of Applied Research in Memory and Cognition 10 (2021) 444–457
Accepted by the outgoing editorial team

Journal of Applied Research in Memory and Cognition

Empirical Article

Wearable Technology for Automatizing Science-Based Study Strategies: Reinforcing Learning Through Intermittent Smartwatch Prompting

Anne M. Cleary*
Colorado State University, USA

Katherine L. McNeely-White
Colorado State University, USA

Hannah Hausman
University of California, Santa Cruz, USA

Jennifer Dawson, Sally Kuhn, Rebecca M. Osborn, Andrew M. Huebert, and Matthew G. Rhodes
Colorado State University, USA

This study explored smartwatches' potential for implementing and automatizing the use of retrieval as a study tool outside of classroom contexts. Across five experiments, review prompts delivered via a smartwatch after reading scientific passages enhanced retention of factual information from the passages and reduced forgetting after a two-day delay. The delivery format—delivery in the form of testing versus in the form of restudying—mattered to the level of learning benefit shown. Consistent with the testing effect, delivering the information as a test question followed a minute later by its answer was generally more beneficial than delivering it as a mere factual statement for restudy. Whether participants were reading magazines, watching Netflix episodes, or engaging with their own smartphones while receiving the smartwatch prompts made no difference to the beneficial effect of the smartwatch prompting on retention. Thus, smartwatch prompts can be applied strategically to automatize outside-of-the-classroom learning reinforcement.

Keywords: Wearable technology, Smartwatch prompting, Testing effect, Reinforcing learning, Preventing forgetting

This document is copyrighted by the American Psychological Association or one of its allied publishers.
Content may be shared at no cost, but any requests to reuse this content in part or whole must go through the American Psychological Association.

Author Note

This research was supported by a grant from the Center for Analytics on Learning and Teaching (C-ALT) at Colorado State University to Anne M. Cleary, which enabled the purchase of four Apple Watches and accompanying iPhones for use in this research.

The data from Experiment 1 were presented at the 2018 APA Technology, Mind, and Society conference in Washington, DC. The data from Experiment 2 were presented at the 2019 Meeting of the Rocky Mountain Psychological Association in Denver, CO.

Some of the data from Experiment 2 were collected by Jennifer Dawson during the summer of 2018 as part of her summer REU (Research Experiences for Undergraduates) project within the Bridges to Baccalaureate program at Colorado State University (NIH 2018 R25 GM).

The data for Experiment 5 were collected by Sally Kuhn for her undergraduate thesis project at Colorado State University during the spring of 2019.

The data and materials from the present study are available on the Open Science Framework at the following link: https://osf.io/yf3s5/.

* Correspondence concerning this article should be addressed to Anne M. Cleary, Department of Psychology, Colorado State University, Fort Collins, CO, United States. Contact: Anne.Cleary@colostate.edu (A.M.C.).

General Audience Summary

Despite an abundance of scientific research demonstrating principles that enhance learning, well-established strategies fail to be widely implemented in the United States. The present study demonstrates that smartwatches can be used to deliver information to a person to boost retention of information from scientific passages, and that smartwatch prompt delivery in the form of testing with feedback confers advantages over and above delivery in the form of simply restudying the information. By making it extremely easy for people to receive automatized reminders of learned information in the form of testing, smartwatches have great promise as a tool that can boost student learning in real-world contexts outside of the classroom.

This document is copyrighted by the American Psychological Association or one of its allied publishers.
Content may be shared at no cost, but any requests to reuse this content in part or whole must go through the American Psychological Association.

The Underutilization of Principles From the Science of Learning in Education

Despite an abundance of research on factors that enhance and optimize learning, educational practice has lagged behind the science (e.g., Dempster, 1988; Carpenter, Cepeda, Rohrer, Kang, & Pashler, 2012; Roediger & Karpicke, 2018; Weinstein, 2018a, 2018b; Willingham, 2018). Dempster (1988) articulated this concern over three decades ago with regard to the principle of spacing. Today, more than 30 years later, the failure to apply spacing in education persists (Weinstein, 2018a, 2018b). Another principle—testing—has faced the same fate. As Roediger and Karpicke (2018) note, the testing effect—the finding that being tested on information leads to better learning than restudying it—was first reported in research over a century ago. Although there has been a recent resurgence of interest among learning scientists, testing remains underutilized as a learning (rather than assessment) tool in educational settings (e.g., see McDaniel, Anderson, Derbish, & Morrisette, 2007). In general, principles gleaned from cognitive psychology as a whole have not been well-translated into educational settings (Willingham, 2018) and many misconceptions about learning prevail (e.g., Anthenien, DeLozier, Neighbors, & Rhodes, 2018; Blasiman, Dunlosky, & Rawson, 2017; Hartwig & Dunlosky, 2012; McCabe, 2011; Morehead, Rhodes, & DeLozier, 2016; Yan, Bjork, & Bjork, 2016).

Possible Contributing Factors to the Underutilization of Effective Learning Strategies

One potential reason why evidence-based learning strategies are underutilized in educational settings is likely that learners may often fail to fully appreciate strategies that best support learning. For example, studies suggest that students think that they learn more from massing than spacing information when the opposite is true (e.g., Kornell, 2009; Kornell & Bjork, 2008; see also King, Zechmeister, & Shaughnessy, 1980). In addition, people tend to think that they learn more by restudying information than being tested on it even though the opposite is true (e.g., Kornell & Son, 2009; Roediger & Karpicke, 2006; Roediger & Karpicke, 2018). Such data indicating misplaced confidence in less effective learning strategies suggest that students will be unlikely to discover the beneficial relationship between effective strategies and learning on their own (e.g., Yan et al., 2016), and survey research on college students' study habits and beliefs is broadly consistent with this notion (e.g., Anthenien et al., 2018; Blasiman et al., 2017; Hartwig & Dunlosky, 2012; McCabe, 2011; Morehead et al., 2016).

Given that students are unlikely to discover the relationship between effective learning strategies and learning outcomes on their own, one strategy put forward by Rhodes, Cleary, and DeLosh (2020) is to explicitly teach students about them, and about the potential disconnect between their feelings and impressions while learning and their actual learning outcomes. Unfortunately, simply teaching students about the disconnect may not be sufficient to elicit the level of behavior change required to implement effective strategies among their study habits (e.g., McDaniel & Einstein, 2020; see also Yan et al., 2016, on the difficulty of mending misconceptions about learning).

Research on behavioral nudges suggests that creating environments that favor a desired behavior (by making it very easy to engage in that behavior) is more effective at eliciting a desired behavior than expecting people to change their own behavior themselves (e.g., Johnson & Goldstein, 2003; Thaler & Sunstein, 2009); thus, expecting people to change their learning habits on their own may not be feasible. As Thaler and Sunstein (2009) suggest, learning about beneficial changes that can be made for self-improvement, even when the prospect excites people, does not typically lead to the required behavior change, particularly if the behavior requires taking action over inaction, or deviating from routine tasks. For example, simply teaching people through educational seminars about how to save more money for retirement does not usually lead to follow-through on implementing what was learned. Thaler and Sunstein (p. 114) report a study by Choi, Laibson, Madrian, and Metrick (2002) that showed that although every person attending an educational seminar on saving more money reported wanting to save more money, only 14% followed through in enrolling for the savings plan being taught about in the educational seminar. Similarly, engaging in effective learning strategies like spacing and testing can require a substantial deviation from a student's existing routines and habits; thus, wanting to engage in effective learning strategies may not be enough to prompt the behavior change necessary to implement them.

As an alternative to simply educating people, Thaler and Sunstein (2009) recommend minimizing the level of required action or behavior change on the part of the people who stand to benefit from that action or behavior change (but see Hertwig & Grüne-Yanoff, 2017, for an alternative perspective). Thaler

and Sunstein suggest that automatization of the circumstances that will most likely benefit people (e.g., automatically enrolling new employees in a 401k) increases the likelihood that people will actually receive that benefit. In the present study, we explored the possibility that a wearable device—specifically, a smartwatch—could effectively automatize some science-based strategies for boosting learning and retention outside the classroom.

How Smartwatches Might Automatize Science-Based Learning Reinforcement Strategies

Smartwatches not only allow for distributed prompting of information over time in a way that can be largely automated, but also have the potential to automate the testing effect through distributed prompting of questions followed by feedback. Specifically, smartwatches can use auditory and haptic alerts to prompt the wearer to read text on the watch face. Such alerts typically happen in the context of upcoming calendar items or text messages. Smartwatches can be programmed by the wearer, or by an app, to signal the wearer to view information on previously learned material at pre-specified times. In this way, previously learned information can potentially be reinforced through distributed delivery over time via the watch, and the testing effect can be implemented by delivering the prompts via questions that encourage attempted retrieval of the answer. Such prompts can be followed by the answer in order to incorporate the benefits of feedback in eliciting the testing effect (Rowland, 2014).

Although a similar implementation is at least theoretically feasible on a smartphone instead of a watch, our reasoning behind testing the implementation on a watch instead of a phone was twofold. First, from a practical standpoint, unless a phone is right in front of a person, that individual may not notice a text message or alert arriving in real time; it is not until the person looks at the phone that the message is seen. The smartwatch is potentially more effective at delivering content that will be seen at the pre-specified time periods, because every message is felt on the wrist when it comes in, and the phone does not have to be in front of the person for this to happen. Second, from an experimental design standpoint, if the person does not click on and dismiss a text message or alert as soon as it arrives on a phone, the messages build up on the phone's screen and then need to be scrolled through manually. Therefore, we would not be able to control how many times participants saw a prompt or for how long they saw the prompt. Thus, for practical and experimental purposes, we favored smartwatches for the present study.

Although it may seem obvious that delivering reinforcers of previously learned information via smartwatch prompting would benefit learning and memory relative to not prompting at all, there are reasons why such prompting might fail to work as intended. Indeed, in most studies of factors that enhance learning, participants are exclusively devoted to the task of learning during the manipulations (e.g., see Rowland, 2014, for a review). The learners' attention is not usually divided such that they would be fully engaged in some other, completely irrelevant task while a smartwatch occasionally distracts them from that task with a reminder of previously learned information. In fact, divided attention and distraction are thought to be detrimental to learning (e.g., Anderson, Craik, & Naveh-Benjamin, 1998; Craik, Govoni, Naveh-Benjamin, & Anderson, 1996; Gaspelin, Ruthruff, & Pashler, 2013; Rhodes et al., 2020; but see Spataro, Mulligan, & Rossi-Arnaud, 2013). Therefore, students may not benefit from smartwatch reinforcement prompts if their attention is divided between the watch prompts and another task (e.g., walking to class, talking to a friend, watching a movie). Moreover, one of the many generalizable principles of learning is that it is beneficial to engage in deep, meaningful processing of the information at hand (e.g., Rhodes et al., 2020). From this perspective, receiving a secondary interruption to a primary task via a smartwatch, especially while fully engaged in another primary task, might even be viewed as antithetical to effective learning. The secondary interruption (i.e., the course content delivered via the smartwatch prompt) may be processed shallowly and thus not encoded well.

Thus, it is not immediately clear that being prompted with smartwatch reminders of previously learned information while engaged in a completely different, irrelevant task would actually enhance learning. It is conceivable that the information presented on the watch will not be fully attended to and thus would not significantly benefit memory. Therefore, an empirical study is needed to assess the utility and feasibility of using smartwatches to automatize learning reinforcements.

The Present Study

The present study sought to investigate the hypothesis that smartwatch delivery of reminders of previously learned information can reinforce that previously learned information. Experiment 1 served as an initial proof of concept that learning reinforcers, in the form of smartwatch prompts, would increase retention of factual information. Subsequent experiments tested the hypothesis that delivery via testing followed by feedback would lead to better retention than delivery only involving restudying the information, and this hypothesis was tested under various conditions including different primary tasks and different retention intervals.

Because it was unclear a priori whether delivering distributed prompts about previously studied information via a smartwatch while engaged in another task would benefit learning at all, we attempted to maximize the likelihood of obtaining a benefit of smartwatch-delivered learning reinforcement in Experiment 1 by solely using testing followed by feedback as the delivery method. This method was chosen because testing followed by feedback leads to a greater learning benefit than simply restudying information or being tested without feedback (Rowland, 2014).

In Experiment 1, participants read four passages (about evolution, ice ages, food allergies, volcanoes) for half an hour. They then read magazines for 50 min while wearing a smartwatch. The smartwatch buzzed the wearer's wrist every five minutes with a question pertaining to the earlier-read passages. One minute following the question, the smartwatch buzzed with the answer, providing the participant with feedback. Participants all took

This document is copyrighted by the American Psychological Association or one of its allied publishers.
Content may be shared at no cost, but any requests to reuse this content in part or whole must go through the American Psychological Association.

the same final multiple-choice test in the final test phase. Experiments 2 through 5 explored the format of smartwatch prompting (testing v.s. restudying), the type of primary task engaged in while prompting occurred (reading magazines, viewing one's own device, watching Netflix), and longer delays between the smartwatch reinforcement and the final test.

Experiment 1

Method

Participants. Participants were 12 Colorado State University undergraduates who participated in exchange for credit toward an introductory-level course. This first experiment was primarily exploratory and an attempt to ascertain the size of the effect of providing reminders via a smartwatch, serving as a pilot experiment for use in applying for funding for four smartwatches, which in turn would enable a level of a scaling-up that would allow more participants to be run in the subsequent experiments. As no research-dedicated smartwatches could be obtained without funding for them, the data for Experiment 1 were collected by running participants one at a time using the personal smartwatch of one of the authors. This limited how many participants could be tested; hence, only 12 were tested. These data were then used to obtain the funding that enabled the subsequent experiments to be run and to be scaled up from Experiment 1.

Design. A three-phase design was used. In the first phase, participants read four scientific passages. In the second phase (an intermediate phase) participants read magazines while wearing a smartwatch that periodically prompted them regarding factual information from two of the four passages that were studied in Phase 1. In the third and final phase, participants received a paper-and-pencil test on the material from all four passages that had been studied in Phase 1. The design was a within-subjects design. Specifically, whether a particular passage had some of its facts prompted through smartwatch prompts or not prompted at all during the intermediate phase was manipulated within-subjects, and which passages were prompted versus not prompted during the intermediate phase was counterbalanced across participants such that odd-numbered participants received smartwatch prompts for one set of the two studied passages and even-numbered participants received smartwatch prompts for the opposite two studied passages. The primary interest was in whether performance on the paper-and-pencil test in Phase 3 would be better for information that had been prompted in the intermediate phase than for information that had not been prompted in that phase. Toward this end, a paired samples *t*-test was used to assess whether the proportion correct was greater for material that had been prompted than for material that had not been prompted in Phase 2.

Materials. The materials were taken from among those used by Hausman and Rhodes (2018), which had been taken from Thiede, Wiley, and Griffin (2011). These materials included expository passages on various topics and multiple-choice test questions pertaining to those passages. The Thiede et al. passages have an average Flesch-Kincaid readability score of 11.8 (Thiede et al., 2011, pp. 266–267). The passages used in the present study were on the following four topics: volcanoes (1080 words), the ice age (1053 words), evolution (582 words), and food allergies (807 words). As in Hausman and Rhodes' study, five factual multiple-choice questions per passage were used to assess learning and memory for the information presented in the passage. The watch prompt questions and their answers in the intermediate phase of the experiment came from the factual multiple-choice test materials (i.e., the final test questions and their answers were identical to those prompted for in the intermediate phase). When time allowed, in Experiments 1, 2, 3, and 5, participants also took an inference test as the last task in the experiment (five inference questions per passage) that was used in Hausman and Rhodes' study. Whereas the factual questions pertained to pieces of information actually stated in the passage text, the inference questions pertained to information not explicitly stated but that instead needed to be inferred by the reader (Thiede et al., 2011).

Procedure. Participants were given 30 min to read through the four passages (volcanoes, the ice age, evolution, and food allergies). They could read through the passages in any order they wished and could go back and review them as they wished, but had to stop after 30 min. For example, if the experiment began at 9:00am, the participant would have until 9:30am to read through the four scientific passages. We did not collect data on how much time each participant spent per passage nor on how many or which participants chose to re-read passages versus not. As the passages were counterbalanced across the experimental conditions, any encoding variability would not vary systematically and would thus not be confounded with the experimental conditions.

At the end of the 30-min reading period, the participant was fitted with a Series 1 edition Apple Watch and then given several printed magazines (of varying genres and options including issues of *National Geographic*, *Time*, *People*, and many others) to read through for the next 50 min. For example, if the participant had begun reading the scientific passages at 9:00am, the watch was fitted shortly after 9:30am after which the person could begin reading and sifting through magazines while awaiting watch prompts. A 5-min buffer was used to allow enough time to fasten the watch to the person's wrist before the first prompt. For example, if the participant had read the four passages from 9:00am until 9:30am and then was fitted with the watch at 9:30am, the watch would buzz for the first time at 9:35am with the first question.

During that 50-min intermediate period of magazine-reading, the watch would buzz the wearer's wrist periodically (on a time schedule) with a question pertaining to one of the earlier-read passages; one minute following the appearance of that question, the watch would buzz again with the question's answer appearing on the watch face. Then, four minutes following the appearance of that answer, the watch would buzz again with the next question, followed one minute later by its answer. For example, if the participant was prompted with the first question (e.g., "How many plates make up the earth's crust?") at 9:35am, then at 9:36am, the watch would buzz again with the answer (e.g., "Answer: 12"). Then, at 9:40am, the watch would buzz again with the next question. Then, at 9:41am, the watch would buzz with that question's answer. The

This document is copyrighted by the American Psychological Association or one of its allied publishers.
Content may be shared at no cost, but any requests to reuse this content in part or whole must go through the American Psychological Association.

prompting of the 10 questions (five questions from each of two of the four studied passages) and their answers continued on this timing schedule for the entire 50-min intermediate phase.

Participants were instructed that when the watch buzzed, they were to turn their wrist toward their head so that the watch face would display for them to show the prompt. They were instructed that once the watch face display was visible, to simply read the question and try to answer it silently in their own mind, then go back to what they were doing and await the next prompt, at which point they should do the same (turn the wrist to activate the display on the watch face). A minute later, they were prompted with the question's answer. Participants were instructed to look at the answer at that time, then go back to what they were doing. An experimenter sat in the room with the participant to ensure that the participant was viewing the watch face display. For example, if the participant kept the watch-wearing wrist still on the table and attempted to merely glance over the watch face without actually turning his or her wrist, the watch face display would not activate (because it activates through wrist-turning). Therefore, the experimenter observed to ensure that the participant understood how to activate the watch display once the watch buzzed and guided the participant through how to do it if the participant initially had trouble.

At the end of the 50-min magazine-reading phase, the watch was removed from the participant's wrist and the participant was then given a paper and pencil multiple-choice test on the earlier-read passages. The factual test contained 20 questions altogether, with five appearing from each of the four different earlier-read passages (volcanoes, evolution, the ice age, and food allergies). The participant was to circle each correct answer with a pencil. If time permitted, the participant was then given the multiple-choice inference test from Hausman and Rhodes (2018), also in paper-and-pencil format, and also with the requirement to circle each correct answer with a pencil.

Results

In addition to using traditional null hypothesis significance testing (NHST), we report the results of Bayesian analyses. Unlike NHST, Bayesian analyses allow us to accept the null hypothesis and not merely fail to reject it (e.g., Kruschke, 2013). Therefore, we report Bayes factors (*BF*s), which quantify the strength of the evidence for the alternative (BF_{10}) and the null hypothesis (BF_{01}) and can be considered weak ($1 < \text{BF} \leq 3$), positive ($3 < \text{BF} \leq 20$), strong ($20 < \text{BF} \leq 150$), or very strong evidence ($\text{BF} > 150$; Wagenmakers, 2007). All Bayes factors were calculated with JASP using the JZS prior because it requires the fewest prior assumptions about the range of the true effect size (Rouder, Speckman, Sun, Morey, & Iverson, 2009).

A significant benefit of smartwatch delivery of questions and answers on final factual test performance was found. A paired-samples *t*-test on the proportion correct for each section on the test revealed better performance among smartwatch prompted items relative to items that were not prompted by the smartwatch, $t(11) = 6.31$, $SE = .06$, $p < .001$, $d = 1.82$, $BF_{10} = 465.99$ (see Table 1 for descriptive statistics). A smartwatch prompting advantage was shown in 11 out of the 12 participants. Finally, because only nine out of the 12 participants completed the inference test in Experiment 1, these data were not examined for Experiment 1.

Table 1
Mean Proportion Correct on the Factual Multiple-Choice Test for Experiments 1 and 2

	Smartwatch prompt		No smartwatch prompt	
Experiment/Condition	*M*	*SD*	*M*	*SD*
Experiment 1 (testing prompts)	.92	.14	.53	.17
Experiment 2 (testing prompts)	.88	.14	.59	.18
Experiment 2 (restudy prompts)	.78	.16	.54	.16

Experiment 2

Experiment 1 served as a proof of concept in showing that, rather than simply being a shallowly processed interruption while a person engaged in another primary task, spaced prompting via a smartwatch with reminders of previously learned material effectively reinforced what was learned and lead to better retention of information. Experiment 2 examined methods of delivering smartwatch reminders. Specifically, does prompting with test questions followed soon after by their answers confer a greater advantage on learning and retention than merely presenting the fact for restudy as a statement? Prior work on the testing effect would suggest that having to retrieve the answer to a question and then receive feedback on it should lead to better learning than merely reading about the answer as a form of restudying it (Rowland, 2014). In Experiment 2, we examined whether this same pattern would apply to spaced smartwatch delivery of previously learned factual information.

Method

Participants. Based on the effect size obtained in Experiment 1 ($d = 1.82$), to achieve a power of 80% and a .05 significance level, only six participants would be needed. However, as Experiment 2 aimed to carry out a between-subjects comparison of the manner of prompting (testing vs. restudying), rather than merely seeking to replicate the finding from Experiment 1, we aimed for a larger sample size, as it was unclear a priori if a testing effect would be found, and if so, what its magnitude would be. Therefore, we aimed to run 60 participants with 30 in each between-subjects condition using participant sign-up estimations and no-show rate estimations to determine a stop date based on this goal. By the stop date, we ended up with 64 participants, who were Colorado State University undergraduates who participated in exchange for credit toward an introductory-level course. The participants were randomly assigned to either a testing condition (smartwatch delivery of test questions each followed one minute later by their answers) or a restudy condition (smartwatch delivery of a restatement of the fact). Four participants were excluded from the analyses due to either being a non-native English speaker, having prior experience with the

This document is copyrighted by the American Psychological Association or one of its allied publishers.
Content may be shared at no cost, but any requests to reuse this content in part or whole must go through the American Psychological Association.

This document is copyrighted by the American Psychological Association or one of its allied publishers.
Content may be shared at no cost, but any requests to reuse this content in part or whole must go through the American Psychological Association.

experiment materials, or the watch not providing prompts due to a technical error. Thirty-one participants ended up in the restudy condition and 29 ended up in the testing condition.

Design. A three-phase design like that used in Experiment 1 was again used (Phase 1: Encoding phase of reading four scientific passages; Phase 2: Intermediate phase of smartwatch prompting from two of the four earlier-read passages during magazine-reading; Phase 3: Final paper-and-pencil test). However, including the type of prompting resulted in a mixed-factor design, whereby smartwatch prompt condition was a within-subjects condition (in which two of the four scientific passages from study were prompted in the intermediate phase and two were not prompted) and prompt type was a between-subjects condition (in which the prompts for the two prompted-for passages were either delivered via testing or via restudying, depending on whether the participant was assigned to the testing condition or to the restudy condition).

Materials. The materials were identical to those used in Experiment 1, with the exception that for the restudy condition the test questions and their corresponding answers were converted to statements of fact. For example, in the testing condition, the test question was "How many plates make up the Earth's crust?" with the answer provided one minute later ("12"). For the restudy condition, participants instead read a statement of the fact (e.g., "12 plates make up the Earth's crust") at the same point in time that the answer would have been delivered in the testing condition.

Procedure. The procedure was identical to that used in Experiment 1, with the exception that participants in the restudy condition received a factual statement every five minutes via the smartwatch prompt, rather than a test question followed by its answer. The timing between the testing and the restudy conditions was equated such that, in the restudy condition, the factual statement occurred at the same point in time that the question's answer would have appeared in the testing condition. For example, if the first question in the testing condition came at 9:35am with its answer followed a minute later at 9:36am, then for a participant assigned to the restudy condition, the factual restatement prompt would come at 9:36am, and the next would come at 9:41am. The within-subjects control condition (prompt versus no prompt) and its corresponding counterbalancing across participants was identical to Experiment 1. As Experiment 2 involved a scaling-up from Experiment 1 through the purchase of four research-dedicated Apple Watches, the experimenter moved between rooms to check on participants rather than sitting one-on-one with a participant.

Results

As Experiment 2 was higher powered than Experiment 1, and more participants were able to take the inference multiple-choice test following the factual multiple-choice test, we were able to analyze the results for the inference test as well as for the factual test. No significant effects of smartwatch prompting on the inference test were found in either Experiment 2, Experiment 3, or Experiment 4 (the inference test was not given in Experiment 5). Therefore, the results of the inference test performance are presented in the Supplementary Materials for Experiments 2, 3, and 4 and will not be discussed further.

Due to errors with the smartwatch during Experiment 2, three participants in the restudy condition did not receive one of the prompts. When computing the proportion of correct answers for these three participants, the question that was not presented via the smartwatch was removed.

A 2×2 Smartwatch Prompt Condition (prompt vs. no prompt) $\times$ Prompt Type (testing vs. restudy) mixed ANOVA performed on the proportion correct on the multiple-choice test revealed a main effect of smartwatch prompt condition, $F(1, 58) = 105.75$, $MSE = .02$, $p < .001$, $\eta_p^2 = .65$, $BF_{10} = 1.87 \times 10^{14}$. As can be seen in Table 1, performance on the multiple-choice test was significantly greater for content for which there had been intermittent smartwatch prompts given during the magazine-reading phase of the experiment, regardless of the delivery method. There was also a main effect of prompt type, $F(1, 58) = 6.79$, $MSE = .03$, $p = .01$, $\eta_p^2 = .11$, $BF_{10} = 1.35$. As shown in Table 1, performance was greater, overall, among participants for whom the prompt delivery method was testing than among those for whom the prompt delivery method was restudying.

Given that no interaction was found, $F(1, 58) = 1.19$, $MSE = .02$, $p = .28$, $\eta_p^2 = .02$, $BF_{01} = 2.15$, it might seem possible that testing via the smartwatch led to an advantage not only for the tested information itself, but also for untested information pertaining to the other passages read by the participants in the testing condition. If so, this would represent a remarkable benefit of testing—that it might extend to the untested information too (cf. Chan, McDermott, & Roediger, 2006). To investigate this more closely, we followed-up with *t*-tests using a conservative Bonferroni corrected alpha level of .008. An independent-samples *t*-test revealed no significant difference between the control conditions from the two comparison groups, $t(58) = 1.26$, $SE = 0.04$, $p = .21$, $d = .32$, $BF_{01} = 1.97$, suggesting that the main effect of prompt type was carried largely by the experimental conditions (i.e., the difference between testing vs. restudying information). Indeed, an independent-samples *t*-test revealed a significant difference between the testing group and the restudy group, $t(58) = 2.90$, $SE = .04$, $p = .005$, $d = .75$, $BF_{10} = 7.84$.

Although providing test questions as prompts led to the highest levels of retention, restudying the information via smartwatch prompting was still beneficial to later memory. A paired-samples *t*-test confirmed that restudying information via the smartwatch, relative to items from which there was no prompt, led to a significant benefit to retention, $t(30) = 6.41$, $SE = .04$, $p < .001$, $d = 1.52$, $BF_{10} = 3.73 \times 10^4$. The same was true of the testing condition relative to its within-subjects control comparison, $t(28) = 8.20$, $SE = .04$, $p < .001$, $d = 1.82$, $BF_{10} = 1.97 \times 10^6$.

The results of Experiment 2 suggest that reminding via smartwatch prompting enhances retention of factual information for a later test on earlier-read scientific passages. In addition, prompting may be optimal when elicited via test questions rather than re-presenting the fact.

This document is copyrighted by the American Psychological Association or one of its allied publishers.
Content may be shared at no cost, but any requests to reuse this content in part or whole must go through the American Psychological Association.

Experiment 3

Thus far, the multiple-choice test at the end of the experiment occurred approximately an hour after the study phase. Experiment 3 investigated if the benefits of the smartwatch prompting during the magazine reading phase would persist across a more substantial delay—2 days—before the multiple-choice test. Toward this end, Experiment 3 employed the same general procedure as Experiment 2, but half of the participants received the multiple-choice test after a delay of two days, and smartwatch prompt type (testing vs. restudying) was manipulated as a within-subjects, rather than a between-subjects, variable.

Method

Participants. In Experiment 3 we aimed to run the same number that we aimed to run in Experiment 3 (60 participants) using a cut-off date based on participant sign-up estimation and no-show estimation. By the cut-off date, we had gotten 58 participants. All were Colorado State University undergraduates who participated in exchange for credit toward an introductory-level course. Participants were randomly assigned to either an immediate or a delayed multiple-choice test condition for the final test of their learning. Due to errors with the watch notification, two participants were lost from the immediate condition. From the delay condition, four participants were lost either due to watch errors or failure to complete the final test within the specified time window of two days. Of the remaining 52 participants, 26 participants were assigned to the immediate condition and 26 to the delay condition.

Design. The same overall three-phase procedure was used as in the previous experiments (Phase 1: Reading of four scientific passages; Phase 2: Reading magazines while being prompted periodically with the smartwatch with information from two of the four passages; Phase 3: A final multiple-choice test administered via Qualtrics). A mixed-factor design was used with delay condition (immediate vs. delayed test) as a between-subjects variable and smartwatch prompt condition (control vs. restudying vs. testing) as a within-subjects variable (where the two unprompted-for passages were the control passages, one of the prompted-for passages had the information prompted for via restudying, and one of the prompted-for passages had the information prompted for via testing followed by feedback).

Materials. The materials were identical to those used in Experiment 2.

Procedure. The procedure was identical to that used in Experiment 2 with the following exceptions. First, the delivery method of prompting via the smartwatch was manipulated within-subjects instead of between-subjects. The control condition of no prompting was also included as a within-subjects comparison, so that the baseline level of forgetting across the delay of two days could be assessed. Participants read the same four scientific passages during the learning phase as in Experiments 1 and 2. During the magazine-reading phase, smartwatch prompts were given for two of the passages. For one of the two prompted-for-passages, the prompts were given in the form of test questions each followed one minute later by their answer whereas for the other of the two passages, the smartwatch prompts were given in the form of statements for restudy, as in Experiment 2. No prompts were given for the remaining two passages (the control passages). Which passages were used for which condition (smartwatch prompting via testing, smartwatch prompting via restudy, or control condition with no prompt) was counterbalanced across participants through random assignment to versions of the experiment.

Second, all participants received the final test via Qualtrics. Third, participants in the delay condition did not receive the multiple-choice test until two days following the magazine reading phase. Participants assigned to this condition left the experiment after the 50 min of smartwatch prompting while reading magazines. Forty-eight hours later, these participants received an email with a link to the multiple-choice test. By following the link during that same day, these participants could take the multiple-choice test online via Qualtrics.

Results

The means and standard deviations from Experiment 3 are presented in Table 2. A 2×3 Delay Condition (immediate test vs. delayed test) $\times$ Smartwatch Prompt Condition (testing vs. restudy vs. no prompt) mixed-measures ANOVA revealed a main effect of smartwatch prompt condition, $F(2, 100) = 27.98$, $MSE = .03$, $p < .001$, $\eta_p^2 = .36$, $BF_{10} = 8.07 \times 10^7$. There was also a significant main effect of Delay Condition, $F(1, 50) = 11.15$, $MSE = .06$, $p = .002$, $\eta_p^2 = .18$, $BF_{10} = 9.56$, suggesting an overall decline in memory from the initial learning. The interaction was not significant, $F(2, 100) = 1.84$, $MSE = .03$, $p = .16$, $\eta_p^2 = .04$, $BF_{01} = 2.16$.

Focusing on the immediate condition, a repeated-measures ANOVA revealed a significant overall effect of smartwatch

Table 2
Mean Proportion Correct on the Multiple-Choice Test Immediately Versus Across a Two-Day Delay

	Smartwatch prompt condition											
	Experiment 3						Experiment 4					
	Test prompts		Restudy prompts		No prompts		Test prompts		Restudy prompts		No prompts	
Delay condition	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
Immediate final test	.89	.13	.76	.21	.59	.20	.91	.15	.79	.21	.62	.18
Delayed final test	.68	.25	.69	.25	.47	.19	.75	.25	.59	.26	.49	.15

This document is copyrighted by the American Psychological Association or one of its allied publishers.
Content may be shared at no cost, but any requests to reuse this content in part or whole must go through the American Psychological Association.

prompt condition (testing vs. restudy vs. no prompt) on multiple-choice test performance, $F(2, 50) = 22.56$, $MSE = .03$, $p < .001$, $\eta_p^2 = .47$, $BF_{10} = 5.89 \times 10^5$. As we had a priori, directional hypotheses regarding the anticipated patterns, paired-samples *t*-tests, rather than post hoc analyses were used. Paired-samples *t*-tests revealed that multiple-choice test performance when the test occurred immediately following the magazine reading phase was higher in the testing condition than in the restudy condition, $t(25) = 2.86$, $SE = .05$, $p = .008$, $d = .74$, $BF_{10} = 5.49$. Performance was also higher in the testing condition than in the control condition for which there were no smartwatch prompts delivered, $t(25) = 8.35$, $SE = .04$, $p < .001$, $d = 1.71$, $BF_{10} = 1.20 \times 10^6$. Smartwatch prompting via restudying also conferred an advantage over no prompting at all, $t(25) = 3.29$, $SE = .05$, $p = .003$, $d = .82$, $BF_{10} = 13.40$.

As reported previously, performance was markedly lower after the two-day delay. However, it is clear that both types of smartwatch prompting led to significantly less forgetting over time than the condition that did not receive prompts and serves as a baseline measure of forgetting. A repeated-measures ANOVA for just the delay condition revealed a significant overall effect of smartwatch prompt condition (testing vs. restudy vs. no prompt) on multiple-choice test performance, $F(2, 50) = 10.20$, $MSE = .04$, $p < .001$, $\eta_p^2 = .29$, $BF_{10} = 270.84$. Paired-samples *t*-tests confirmed that participants demonstrated better performance on the multiple-choice test when prompted via testing compared to not being prompted, $t(25) = 3.72$, $SE = .06$, $p = .001$, $d = 1.00$, $BF_{10} = 33.82$, and also when prompted via restudying compared to not being prompted, $t(25) = 4.50$, $SE = .05$, $p < .001$, $d = .95$, $BF_{10} = 203.30$. In short, both forms of prompting led to significant resistance to forgetting across the two-day period. Thus, even a mere 50-min reinforcement period following an initial learning period, whereby smartwatch prompts are delivered for just 50-min to reinforce what was learned, can lead to significant resistance to forgetting over a delay of two days.

Notably, the type of prompting—whether the prompts came in the form of test questions or in the form of factual statements for restudying—did not bear on forgetting. A paired-samples *t*-test revealed that, after a delay of two days, performance on the multiple-choice test was roughly equivalent for the two types of prompting after the two day delay, $t(25) = 0.13$, $SE = .06$, $p = .90$, $d = .03$, $BF_{01} = 4.78$. Thus, although the form of smartwatch prompting made a difference when the test occurred immediately after the prompting session, both forms of smartwatch prompting led to the same overall benefit to resisting forgetting across a delay of two days, in contrast to prior reports of enhanced effects of testing across delays (e.g., Roediger & Karpicke, 2006; see also Kornell, Bjork, & Garcia, 2011). Given a single experiment, we treat the finding as tentative and suggest that it warrants additional replication, which we attempt to do in Experiment 4.

Experiment 4

The data presented across the three experiments thus far demonstrate robust effects of smartwatch prompting on retention of information. Experiment 4 served to attempt to replicate Experiment 3's lack of a testing effect after a two-day delay using a higher-powered design for comparing the immediate versus delay conditions, while at the same time examining the effects of watch-prompting on yet another type of potentially engaging primary task: watching Netflix.

Method

Participants. To achieve more power than was achieved in Experiment 3, we aimed to obtain roughly twice as many participants in the immediate and the delay conditions of Experiment 4 as were obtained in Experiment 3. As there were 26 participants in each of these two conditions in Experiment 3, we aimed to obtain 50 per condition in Experiment 4. However, by the cut-off date, we had obtained 89 participants. These participants were all undergraduates from Colorado State University who completed the experiment in exchange for course credit and were randomly assigned to either the Immediate Testing or the Delay Testing condition. Forty-two participants completed the Immediate Testing condition; however, two of them were lost from analyses due to either failing to complete the experiment or technical problems, leaving a total of 40 participants in this condition. There were 47 participants who completed the Delay Testing condition; however, seven were lost from analyses due to either technical errors with the smartwatches or not completing the final test, resulting in 40 participants being left in the Delay Testing condition. Thus, there were 40 participants in the Immediate Testing condition and 40 participants in the Delay Testing condition.

Design. The same three-phase procedure that had been used in Experiment 3 was used in Experiment 4 with the exception that Phase 2 now involved the primary task of watching Netflix instead of reading magazines (Phase 1: Reading of four scientific passages; Phase 2: Watching Netflix while being prompted periodically with the smartwatch; Phase 3: A final multiple-choice test administered via Qualtrics). As in Experiment 3, a mixed-factor design was used with delay condition (immediate vs. delayed test) as a between-subjects variable and smartwatch prompt condition (control vs. restudying vs. testing) as a within-subjects variable.

Materials. The materials were the same as those used in Experiment 3, with the exception that the primary task was watching Netflix episodes instead of reading magazines. Participants were allowed to choose from a list of six series, including *The Office*, *New Girl*, *Parks and Recreation*, *Gilmore Girls*, *That 70s Show*, and *Friends*.

Procedure. The procedure was identical to that used in Experiment 3, but with the exception that participants were tasked with watching Netflix during the smartwatch prompting phase. They were allowed to select any of the six series listed in the Materials section and could freely switch between shows. Additionally, at the end of the experiment, participants rated how likely it would be for them to use a smartwatch in addition to their typical study habits on a scale of 1 (*not at all likely*) to 10 (*extremelylikely*). Participants were also asked the open question: "Do you foresee any barriers to using a smartwatch to study? If so, please list them."

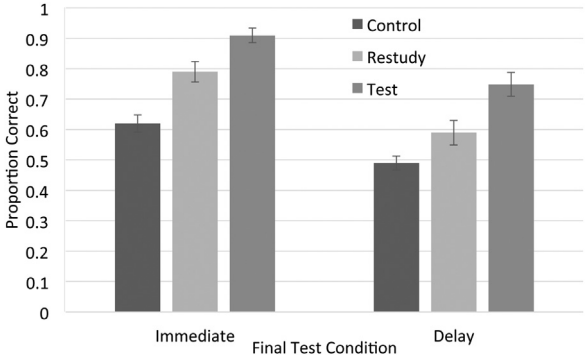

Figure 1. Average final test performance as a function of watch prompt condition and testing delay in Experiment 4, where participants watched Netflix while being prompted via the watch.

This document is copyrighted by the American Psychological Association or one of its allied publishers.
Content may be shared at no cost, but any requests to reuse this content in part or whole must go through the American Psychological Association.

Results

A 2 × 3 Delay Condition (immediate test vs. delayed test) × Smartwatch Prompt Condition (testing vs. restudy vs. no prompt) mixed-measures ANOVA revealed a significant main effect of smartwatch prompt condition on factual multiple-choice test performance, $F(2, 156) = 48.36$, $MSE = .03$, $p < .001$, $\eta_p^2 = .38$, $BF_{10} = 2.60 \times 10^{14}$ (see Figure 1 and the right-hand panel of Table 2). As can be seen in Figure 1, there was also a significant of main effect of delay condition, $F(1, 78) = 26.10$, $MSE = .06$, $p < .001$, $\eta_p^2 = .25$, $BF_{10} = 2.25 \times 10^3$, such that performance overall was higher in the immediate testing condition than the delayed testing condition. The interaction was not significant, $F(2, 156) = .79$, $MSE = .03$, $p = .46$, $\eta_p^2 = .01$, $BF_{10} = .15$.

Starting with the effect of smartwatch prompt condition in the immediate testing condition, a repeated-measures ANOVA performed on multiple-choice test performance revealed a significant effect of smartwatch prompt condition, $F(2, 78) = 29.82$, $MSE = .03$, $p < .001$, $\eta_p^2 = .43$, $BF_{10} = 5.05 \times 10^8$. Paired-samples *t*-tests revealed that multiple-choice test performance was higher in the testing condition than the restudy condition, $t(39) = 3.59$, $SE = .03$, $p = .001$, $d = .65$, $BF_{10} = 33.04$; thus, a testing effect was shown. When comparing performance in the control condition to the testing condition, a significant difference was also found, $t(39) = -8.28$, $SE = .04$, $p < .001$, $d = -1.75$, $BF_{10} = 2.66 \times 10^7$, as well as when comparing the control condition to the restudy condition, $t(39) = -3.87$, $SE = .04$, $p < .001$, $d = -0.87$, $BF_{10} = 69.59$.

For the delayed testing condition, a repeated-measures ANOVA performed on multiple-choice test performance revealed a significant effect of smartwatch prompt condition, $F(2, 78) = 20.15$, $MSE = .03$, $p < .001$, $\eta_p^2 = .34$, $BF_{10} = 2.74 \times 10^5$. Like the immediate testing condition, multiple-choice test performance was higher for items from the testing condition than for items from the restudy condition, $t(39) = -3.95$, $SE = .04$, $p < .001$, $d = -0.63$, $BF_{10} = 86.23$. Thus, unlike in Experiment 3, a testing effect was found in the delayed condition. This suggests that the lack of a testing effect after a two-day delay in Experiment 3 was not due to a difference in testing circumstances (supervised in the lab vs. unsupervised via an email link in the delayed condition), as the testing circumstances were the same in Experiment 4 as in Experiment 3. Participants also demonstrated higher performance on the multiple-choice test when prompted via testing compared to not being prompted, $t(39) = 6.39$, $SE = .04$, $p < .001$, $d = 1.24$, $BF_{10} = 1.02 \times 10^5$, as well as when prompted via restudying compared to not being prompted, $t(39) = 2.34$, $SE = .04$, $p = .02$, $d = 0.47$, $BF_{10} = 1.93$.

Not surprisingly, multiple-choice test performance declined after a two-day delay relative to receiving the test immediately following the intermediate phase of watching Netflix while being prompted with the smartwatch. In addition to the aforementioned main effect of delay condition depicted in Figure 1, multiple-choice test performance among the control conditions (for material that was not prompted via the smartwatch) was higher in the immediate testing condition than in the delayed test condition, $t(78) = 3.54$, $SE = .04$, $p = .001$, $d = 0.78$, $BF_{10} = 42.6$. The same was found for material that was prompted via restudying, with participants in the immediate condition demonstrating higher performance than participants in the delay condition, $t(78) = 3.80$, $SE = .05$, $p < .001$, $d = 0.85$, $BF_{10} = 90.14$. Finally, the same was found when material was prompted via testing: Participants in the Immediate condition exhibited higher performance than participants in the delay condition, $t(64) = 3.53$, $SE = .05$, $p = .001$, $d = 0.78$, $BF_{10} = 41.17$ (note that Levene's test for equality of variances was violated, resulting in varying degrees of freedom). All of this is attributable to general forgetting across the two-day delay relative to being tested immediately after watching Netflix.

Besides extending the generality of our smartwatch prompting findings to the more life-like primary task of Netflix-watching and attempting to determine if a testing effect would be found across the delay condition in a higher-powered experiment, another goal of Experiment 4 was to examine participants' ratings of their likelihood of using a smartwatch to study in addition to their regular study habits. On average, participants provided a rating of 5.84 ($SD = 3.03$) on the scale of 1 – 10, which did not significantly differ between the immediate testing ($M = 6.08$, $SD = 3.06$) and the delayed testing ($M = 5.60$, $SD = 3.02$) conditions, $t(77) = 0.70$, $SE = .68$, $p = .49$. In response to the request to share what obstacles they foresaw with using a smartwatch as a study tool, 48 participants (60%) responded. From these responses, five general themes emerged, with the most common theme being *potentialdistractions* ($N = 20$, 42%). Participants shared that they felt that there could be too many distractions for the smartwatch to be effective (e.g., "The other notifications that you receive on your watch could distract from the studying ability."). Additionally, the next most common theme was *smartwatchinterface* ($N = 14$, 29.2%), with participants sharing that in addition to the watch having a small screen, the time it would take to set up the notifications could be better spent elsewhere. For example, one participant wrote, "As of right now there really isn't any viable smartwatch supported studying apps available, and the use of google calendar reminders can become cumbersome for the watch wearer. . ."

Another common theme was that some participants did not foresee any barriers and that the watch could be *potentiallyhelpful* ($N = 14$, 29.2%), with one participant writing, "No I loved

the idea of a smartwatch because I was able to do something I liked while still being reminded of the facts I needed to learn." The other two general themes that emerged from the responses were *money* ($N = 8$, 16.7%; "Can't afford one.") and *ineffective* ($N = 7$, 14.6%). There were participants who felt that the smartwatch was not an effective study strategy, with one participant sharing, "It does not engage me deep enough for there to be good encoding of information. Paying attention to a smartwatch notification takes only part of my attention, not all of it like traditional studying." Given our findings, however, this could be yet another example of the aforementioned disconnect between student impressions during learning and actual learning outcomes (e.g., Kornell & Son, 2009; Roediger & Karpicke, 2006; Roediger & Karpicke, 2018).

Experiment 5

Experiments 1–4 demonstrated a robust benefit of smartwatch prompting as a means of reinforcing previously learned information and providing resistance to forgetting over time. In these prior experiments, we employed either magazine reading or Netflix watching during the interpolated period before the final test. In Experiment 5, we examined whether the benefit of smartwatch prompting maintains when the primary task is freely interacting with one's own device, and if so, how this effect compares to when the primary task is reading magazines.

Method

Participants. The method used was identical to that used in Experiment 1, for which the effect size was 1.82 (and for which the effect size for the identical condition in Experiment 2 was also 1.82). Based on this effect size, to achieve a power of 80% and a .05 significance level, only six participants would be needed, but because we were seeking to compare two different primary task conditions in the intermediate phase (magazine-reading versus engaging with participants' own devices), we aimed to run 30–40 participants, with 15–20 in each condition. Based on use of a cut-off date for discontinuing running the experiment during a very slow recruitment period, we obtained 34 participants altogether, with 17 in each condition. All 34 were Colorado State University undergraduates, 21 of whom participated in exchange for credit toward an introductory-level course and 13 of whom were compensated with $20 cash. The students who participated in exchange for course credit came from the Psychology Department Research Participation pool whereas those who participated in exchange for pay were recruited from around campus via a recruitment flyer (the limited funds to pay participants were from an Honors Thesis Improvement Grant awarded by the Colorado State University Honors Program to S. Kuhn, and were used to expand participant recruitment beyond the Psychology Department in order to increase sample size).

Participants were randomly assigned to either a magazine reading condition or a smartphone condition in which the participants were free to look at and interact with their own phone, or otherwise use the time however they wished with their own devices. No participants were lost in this experiment.

Design. The same overall three-phase procedure was used as in Experiment 1 (Phase 1: Reading four scientific passages; Phase 2: Reading magazines or engaging with one's own device while being prompted periodically with the smartwatch (with test questions followed one minute later by the answer); Phase 3: An immediate final paper-and-pencil multiple-choice test). A 2×2 mixed-factor design was used with smartwatch prompt condition (prompt vs. no prompt) as a within-subjects variable and interim activity (magazine-reading vs. smartphone-use) as a between-subjects variable.

Materials. The materials were identical to those used in Experiment 1, with the exception that the primary task for the intermediate phase in one condition was personal device usage, rather than magazine reading, and the final test materials consisted only of the factual test. The inference test was not included because a short survey on participants' personal device usage was given instead (the results of this survey are reported in the Supplementary Materials in Figure S1).

Procedure. The procedure was identical to that used in Experiment 1, with the exception that during the intermediate phase (Phase 2), participants in the smartphone condition, rather than being instructed to read magazines, were permitted to use their own smartphones or other devices however they wished during the 50-min smartwatch-prompting period (though magazines were still available and they could choose to spend their time that way; see Supplementary Materials for how many chose to spend some of their time this way).

Results

The means and standard deviations for Experiment 5 are presented in Table 3. A 2×2 Smartwatch Prompt Condition (prompt vs. no prompt) $\times$ Interim Activity (magazine-reading vs. smartphone-use) mixed ANOVA revealed a significant main effect of smartwatch prompt condition, $F(1, 32) = 80.89$, $MSE = .02$, $p < .001$, $\eta_p^2 = .72$, $BF_{10} = 6.46 \times 10^8$. This main effect was such that participants exhibited better performance on the multiple-choice test when the information had been prompted via the smartwatch than when it had not. There was no interaction found ($F < 1.0$, $BF_{01} = 2.52$), nor was there a main effect of interim activity ($F < 1.0$, $BF_{01} = 7.36$).

An independent samples *t*-test revealed no significant difference between being prompted via the smartwatch in the magazine condition versus in the smartphone condition, $t(32) = -0.40$, $SE = .06$, $p = .70$, $d = -0.13$, $BF_{01} = 2.87$. An independent samples *t*-test also revealed no significant differ-

Table 3
Mean Proportion Correct on the Multiple-Choice Test for Experiment 5

	Smartwatch prompt		No smartwatch prompt	
Interim task condition	*M*	*SD*	*M*	*SD*
Magazine-reading	.92	.11	.59	.22
Engaging with Smartphone	.90	.22	.62	.17

This document is copyrighted by the American Psychological Association or one of its allied publishers.
Content may be shared at no cost, but any requests to reuse this content in part or whole must go through the American Psychological Association.

ence between the control conditions for the magazine versus the smartphone interim activities, $t(32) = 0.53$, $SE = .07$, $p = .60$, $d = 0.18$, $BF_{01} = 2.73$. Paired samples t-tests revealed that, within each condition, there was a significant benefit to being prompted with the smartwatch, both in the magazine-reading condition, $t(16) = 7.16$, $SE = .05$, $p < .001$, $d = 1.92$, $BF_{10} = 8.48 \times 10^3$, and in the smartphone condition, $t(16) = 5.60$, $SE = .05$, $p < .001$, $d = 1.36$, $BF_{10} = 640.53$.

Overall, data from Experiment 5 suggest that the type of primary task—the task in which the participant is primarily engaged while being periodically interrupted with the smartwatch prompts—did not have an impact on the level of learning benefit shown. It is conceivable that use of one's own smartphone would be more engaging than reading magazines. If so, the pattern obtained here would suggest that smartwatch reminders of previously learned information should be effective reinforcers of that information even when the watch-wearer is fully engrossed in the primary task at hand while being prompted.

General Discussion

The present study demonstrates the potential utility of smartwatches as a means of reinforcing previously learned information. Across five experiments, performance on a final multiple-choice factual test was better for information that was prompted with a smartwatch during an intermediate phase, regardless of whether the primary task in which participants were engaged was reading magazines (Experiments 1–3), watching Netflix (Experiment 4), or looking at their own devices (Experiment 5). Smartwatch prompting also reduced forgetting across a two-day delay, even when the prompts were in the form of restudying, rather than testing. That said, when the smartwatch prompting occurred in the form of a question (e.g., "How many plates make up the Earth's crust?") followed one minute later with the answer (e.g., "Answer: 12"), performance was better than when the smartwatch prompting occurred in the form of statements of fact for restudy. Experiment 3 was the only experiment in which a testing effect was not found in one of the conditions (the two-day delay condition); however, Experiment 4, which was higher-powered and involved a more realistic primary task (watching Netflix), demonstrated a significant testing effect after the two-day delay.

The present results have practical implications. In no experiment was testing worse than restudying; in fact, in all but one instance, testing led to greater benefits than restudying. Therefore, if one is going to use smartwatch prompts for reinforcing learning and aiding retention, prompts in the form of testing are likely better than prompts in the form of restudying information. Our evidence suggests no harm in using testing instead of restudying, and the potential for greater benefit.

Another practical aspect of the present findings is that the learning reinforcement attained from intermittent smartwatch prompting does not require much overt behavior change on the part of the learner or an instructor in order for the strategy to be implemented. Thus, the method presents a means of potentially automatizing the testing effect, rather than requiring a dramatic change in study or teaching habits in order for learners to reap the benefits of testing outside of the classroom. Their mere convenience and ease of usage with little required behavior change may make them an ideal means of getting students to actually implement distributed testing in their daily lives. Our results suggest that students could be watching Netflix in the evening while reinforcing their learning and strengthening their retention at the same time through intermittent smartwatch prompting.

The present study provides a basis for developing a specialized application for implementing the testing effect via smartwatch prompting. An app could schedule the intermittent prompting across multiple different days, allowing for the added benefits of spaced testing (e.g., Rawson, Dunlosky, & Sciartelli, 2013; Rawson, Vaughn, & Carpenter, 2015; Rawson, Vaughn, Walsh, & Dunlosky, 2018; see also Bahrick, 1979) and getting episodes of sleep in between the smartwatch sessions (cf. Born, Rasch, & Gais, 2006; Wagner, Gais, Haider, Verleger, & Born, 2004). In this way, the spacing of these alerts across days can be automatized, simply occurring at their pre-scheduled times without much behavior change required by the wearer. The alerts can be set to run every evening during down time, such as when the learner watches Netflix or other streamed content. An app could also incorporate feedback from the user, such as by asking after each test question's answer prompt, "Did you get it right? Yes or no?" The app could then select questions for later testing that were gotten wrong.

As a word of caution, we do not recommend scheduling learning reinforcement alerts to occur during activities for which safety is of concern if attention is divided away from the primary task, such as while driving, as evidence suggests that dividing one's attention while driving negatively impacts reaction time for braking (e.g., Strayer & Johnston, 2001). Likewise, it is prudent to not schedule learning reinforcement alerts to occur in situations for which turning to one's watch might be considered by others to be rude, such as when attending a class, interviewing for a job, or public speaking. Again, the present study demonstrates that it is feasible to schedule such alerts to occur during "down time."

The potential benefits of the smartwatch delivery method of automatizing distributed reinforcements of previously learned information is not limited to students attempting to learn coursework. Companies and organizations could potentially capitalize on this method for reinforcing important critical information taught to employees in training sessions. Self-learners trying to learn new vocabulary words or foreign terms could use the smartwatch method as a reinforcer.

Future Research

The present experiments suggest an agenda for future research. For example, future research should examine different prompting intervals and schedules to determine if there are optimal spacing schedules for smartwatch delivery of information, and if the optimal schedule depends on the type of learning involved or the time until the final test (Cepeda, Vul, Rohrer, Wixted, & Pashler, 2008). The ideal spacing schedule, number of prompts within a day, and time span in which the alerts occur

This document is copyrighted by the American Psychological Association or one of its allied publishers.
Content may be shared at no cost, but any requests to reuse this content in part or whole must go through the American Psychological Association.

remain important topics for future research (both in terms of the boost to learning and of learner tolerance of the smartwatch prompts). In the present study, the smartwatch prompts were delivered just minutes apart within an hour; however, smartwatches provide a wide range of possibilities regarding the frequency and spacing of alerts. The sequence of alerts can be scheduled for a particular time of day and in close proximity to one another such that it only takes 10–15 minutes to get through the prompts, or it can be more spread out; future research should examine the ideal timing format.

Regarding *how* to best test using smartwatch prompts, self-generating answers rather than choosing from among multiple-choice options leads to better learning outcomes (e.g., Rowland, 2014). Although the present study's smartwatch test questions were short-answer, other testing formats are possible. Smartwatches can deliver images (which can be used to create multiple-choice options), and can deliver passages of text to scroll through. That said, lengthy text or complex images might display more awkwardly on a smartwatch than on a smartphone, making shorter test questions and/or simple images their ideal prompting format. Still, it is likely advantageous to learning to have the wearer attempt to self-generate an answer rather than to recognize it from among a set of options (Rowland, 2014), and nothing precludes mentally generating longer answers on the part of the wearer. Moreover, Heitmayer and Lahlou (2021) recently showed that only 11% of smartphone interactions are notification-initiated, pointing toward another potential advantage of smartwatches over smartphones for uses requiring frequent notifications, including other forms of education-related nudges (e.g., Motz, Canning, Green, Quick, & Mallon, 2020a; Motz, Mallon, & Quick, 2020b). Therefore, future research should compare smartwatch to smartphone delivery formats. Researchers could also examine encoding manipulations like elaborative interrogation (e.g., Woloshyn, Pressley, & Schneider, 1992) via smartwatch, as well as whether there are metacognitive disconnects between student impressions of their learning benefits and the actual learning benefits from smartwatch prompting.

Finally, future research should aim to extend these findings beyond expository text to inductive learning, skill acquisition, creative problem solving, and courses that require the formation of a deep new understanding. For instance, given the benefits of spacing for problem solving (e.g., Kounios & Beeman, 2015), spaced reminders to think more about a problem or concept might help aid understanding.

Author Statement

A.C. conceived of the overarching idea, collected most of the data for Experiment 1 and analyzed it, applied for and received funding to purchase the smartwatches for the subsequent experiments, oversaw the logistical lab operation for the smartwatch experiments, and took the lead in writing up the present report. K.W. helped with the logistics of setting up and managing multiple smartwatches in the lab for this project, contributing to the design of Experiments 2–4, did the majority of the work on coding and analyzing the data for Experiments 2–4, and wrote the results sections for Experiments 2–4. K.W. also prepared the data for uploading onto the OSF for later public access. H.H. assisted with experimental design for Experiments 1–3, collected some of the data for Experiment 1, and performed the Bayesian analyses reported in the present study as well as contributed to the writing of the manuscript. J.D. collected some of the data for Experiment 2 as part of her summer REU project and contributed to the design of that experiment and some of the analyses. S.K. conceived of the idea for Experiment 4 and contributed to its design; she collected the data for that experiment and performed the data analysis with the help of K.W., including creating Figure 1. B.O. set up the Qualtrics tests to allow for the two-day delay condition used in Experiment 3, and assisted with processing those data from Qualtrics. A.H. helped with the logistical management of setting up smartwatches in the lab on a regular basis and troubleshooting, as well as contributing to editing the manuscript. Finally, M.R. contributed to the design of the experiments starting in the earliest stages of the project, and to the writing of the manuscript by providing substantial edits and suggestions.

Conflict of Interest

The authors declare that they have no conflict of interest.

Online Supplement

Supplementary material related to this article can be found, in the online version, at https://doi.org/10.1016/j.jarmac.2021.01.001.

References

Anderson, N. D., Craik, F. I., & Naveh-Benjamin, M. (1998). The attentional demands of encoding and retrieval in younger and older adults: 1. Evidence from divided attention costs. *Psychology and Aging*, *13*, 405–423.

Anthenien, A. M., DeLozier, S. J., Neighbors, C., & Rhodes, M. G. (2018). College student normative misperceptions of peer study habit use. *Social Psychology of Education*, *21*, 303–322.

Bahrick, H. P. (1979). Maintenance of knowledge: Questions about memory we forgot to ask. *Journal of Experimental Psychology: General*, *108*, 296–308.

Blasiman, R. N., Dunlosky, J., & Rawson, K. A. (2017). The what, how much, and when of study strategies: Comparing intended versus actual study behaviour. *Memory*, *25*, 784–792.

Born, J., Rasch, B., & Gais, S. (2006). Sleep to remember. *The Neuroscientist*, *12*, 410–424.

Carpenter, S. K., Cepeda, N. J., Rohrer, D., Kang, S. H., & Pashler, H. (2012). Using spacing to enhance diverse forms of learning: Review of recent research and implications for instruction. *Educational Psychology Review*, *24*, 369–378.

Cepeda, N. J., Vul, E., Rohrer, D., Wixted, J. T., & Pashler, H. (2008). Spacing effects in learning: A temporal ridgeline of optimal retention. *Psychological Science*, *19*, 1095–1102.

Chan, J. C., McDermott, K. B., & Roediger, H. L., III. (2006). Retrieval-induced facilitation: Initially nontested material can benefit from prior testing of related material. *Journal of Experimental Psychology: General*, *135*, 553–571.

This document is copyrighted by the American Psychological Association or one of its allied publishers.
Content may be shared at no cost, but any requests to reuse this content in part or whole must go through the American Psychological Association.

Choi, J. J., Laibson, D., Madrian, B., & Metrick, A. (2002). Defined contribution pensions: Plan rules, participant decisions, and the path of least resistance. In J. Poterba (Ed.), *Tax policy and the economy* (Vol. 16, pp. 67–113). Cambridge: MIT Press.

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*, 159–180.

Dempster, F. N. (1988). The spacing effect: A case study in the failure to apply the results of psychological research. *American Psychologist*, *43*, 627–634.

Gaspelin, N., Ruthruff, E., & Pashler, H. (2013). Divided attention: An undesirable difficulty in memory retention. *Memory & Cognition*, *41*, 978–988.

Hartwig, M. K., & Dunlosky, J. (2012). Study strategies of college students: Are self-testing and scheduling related to achievement? *Psychonomic Bulletin & Review*, *19*, 126–134.

Hausman, H., & Rhodes, M. G. (2018). When pre-testing fails to enhance learning concepts from reading text. *Journal of Experimental Psychology: Applied*, *24*, 331–346.

Heitmayer, M., & Lahlou, S. (2021). Why are smartphones disruptive? An empirical study of smartphone use in real-life contexts. *Computers in Human Behavior*, *116*, 106637.

Hertwig, R., & Grüne-Yanoff, T. (2017). Nudging and boosting: Steering or empowering good decisions. *Perspectives on Psychological Science*, *12*, 973–986.

Johnson, E. J., & Goldstein, D. (2003). Do defaults save lives? *Science*, *302*, 1338–1339.

King, J. F., Zechmeister, E. B., & Shaughnessy, J. J. (1980). Judgments of knowing: The influence of retrieval practice. *The American Journal of Psychology*, 329–343.

Kornell, N. (2009). Optimizing learning using flashcards: Spacing is more effective than cramming. *Applied Cognitive Psychology*, *23*, 1297–1317.

Kornell, N., & Bjork, R. A. (2008). Learning concepts and categories: Is spacing the "enemy of induction"? *Psychological Science*, *19*, 585–592.

Kornell, N., Bjork, R. A., & Garcia, M. A. (2011). Why tests appear to prevent forgetting: A distribution-based bifurcation model. *Journal of Memory and Language*, *65*, 85–97.

Kornell, N., & Son, L. K. (2009). Learners' choices and beliefs about self-testing. *Memory*, *17*, 493–501.

Kounios, J., & Beeman, M. (2015). *The Eureka Factor: Aha Moments, Creative Insight, and the Brain.* New York, NY: Random House.

Kruschke, J. K. (2013). Bayesian estimation supersedes the t test. *Journal of Experimental Psychology: General*, *142*, 573–603.

McCabe, J. (2011). Metacognitive awareness of learning strategies in undergraduates. *Memory & Cognition*, *39*, 462–476.

McDaniel, M. A., Anderson, J. L., Derbish, M. H., & Morrisette, N. (2007). Testing the testing effect in the classroom. *European Journal of Cognitive Psychology*, *19*, 494–513. https://doi.org/10.1080/09541440701326154

McDaniel, M. A., & Einstein, G. O. (2007). Training learning strategies to promote self-regulation and transfer: The knowledge, belief, commitment, and planning framework. *Perspectives on Psychological Science*, *15*, 494–513.

Morehead, K., Rhodes, M. G., & DeLozier, S. (2016). Instructor and student knowledge of study strategies. *Memory*, *24*, 257–271.

Motz, B., Canning, E., Green, D., Quick, J., & Mallon, M. (2020). *The efficacy of automated praise at facilitating behavior change*. https://doi.org/10.31234/osf.io/7rc3j

Motz, B., Mallon, M., & Quick, J. (2020). *Automated educative nudges to reduce missed assignments in college*. https://doi.org/10.35542/osf.io/u263b

Rawson, K. A., Vaughn, K. E., & Carpenter, S. K. (2015). Does the benefit of testing depend on lag, and if so, why? Evaluating the elaborative retrieval hypothesis. *Memory & Cognition*, *43*, 619–633.

Rawson, K. A., Vaughn, K. E., Walsh, M., & Dunlosky, J. (2018). Investigating and explaining the effects of successive relearning on long-term retention. *Journal of Experimental Psychology: Applied*, *24*, 57–71.

Rawson, K. A., Dunlosky, J., & Sciartelli, S. M. (2013). The power of successive relearning: Improving performance on course exams and long-term retention. *Educational Psychology Review*, *25*, 523–548.

Rhodes, M. G., Cleary, A. M., & DeLosh, E. L. (2020). *A guide to effective studying and learning: Practical strategies from the science of learning*. New York, NY: Oxford University Press.

Roediger, H. L., III, & Karpicke, J. D. (2006). Test-enhanced learning: Taking memory tests improves long-term retention. *Psychological Science*, *17*, 249–255.

Roediger, H. L., III, & Karpicke, J. D. (2018). Reflections on the resurgence of interest in the testing effect. *Perspectives on Psychological Science*, *13*, 236–241.

Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., & Iverson, G. (2009). Bayesian t tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin & Review*, *16*, 225–237.

Rowland, C. A. (2014). The effect of testing versus restudy on retention: A meta-analytic review of the testing effect. *Psychological Bulletin*, *140*, 1432–1463.

Spataro, P., Mulligan, N. W., & Rossi-Arnaud, C. (2013). Divided attention can enhance memory encoding: The attentional boost effect in implicit memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *39*, 1223–1231.

Strayer, D. L., & Johnston, W. A. (2001). Driven to distraction: Dual-task studies of simulated driving and conversing on a cellular telephone. *Psychological Science*, *12*, 462–466.

Thaler, R. H., & Sunstein, C. R. (2009). *Nudge: Improving decisions about health, wealth, and happiness*. London, UK: Penguin Books.

Thiede, K. W., Wiley, J., & Griffin, T. D. (2011). Test expectancy affects metacomprehension accuracy. *British Journal of Educational Psychology*, *81*, 264–273.

Wagner, U., Gais, S., Haider, H., Verleger, R., & Born, J. (2004). Sleep inspires insight. *Nature*, *427*, 352–355.

Wagenmakers, E. J. (2007). A practical solution to the pervasive problems of*p* values. *Psychonomic Bulletin & Review*, *14*, 779–804. https://doi.org/10.3758/BF03194105

Weinstein, Y. (2018a). *The failed spacing effect 30 years later (Part I)*. Blog post at LearningScientists.org. Posted on August 31, 2018. https://www.learningscientists.org/blog/2018/8/31-1.

Weinstein, Y. (2018b). *The failed spacing effect 30 years later (Part 2)*.. Blog post at LearningScientists.org. Posted on September 9, 2018. https://www.learningscientists.org/blog/2018/9/8-1

Willingham, D. (2018). Unlocking the science of how kids think. *Education Next*, *18*, https://www.educationnext.org/unlocking-science-how-kids-think-new-proposal-for-reforming-teacher-education/.

Woloshyn, V. E., Pressley, M., & Schneider, W. (1992). Elaborative interrogation and prior knowledge effects on learning of facts. *Journal of Educational Psychology*, *84*, 115–124.

This document is copyrighted by the American Psychological Association or one of its allied publishers.
Content may be shared at no cost, but any requests to reuse this content in part or whole must go through the American Psychological Association.

Yan, V. X., Bjork, E. L., & Bjork, R. A. (2016). On the difficulty of mending metacognitive illusions: A priori theories, fluency effects, and misattributions of the interleaving benefit. *Journal of Experimental Psychology: General*, *145*, 918–933.

Received 15 September 2020;
Received in revised from 11 January 2021;
Accepted 12 January 2021;
Available online 27 February 2021

This document is copyrighted by the American Psychological Association or one of its allied publishers.
Content may be shared at no cost, but any requests to reuse this content in part or whole must go through the American Psychological Association.