After-Action Reviews and Long-Term Performance: An Experimental Examination in the Context of an Emergency Simulation

Gonzalo J. Muñoz, **Diego A. Cortéz**, **Constanza B. Álvarez**, **Juan A. Raggio**, **Antonia Concha**, **Francisca I. Rojas**, Universidad Adolfo Ibáñez, Santiago, Chile, **Winfred Arthur, Jr.**, Texas A&M University, College Station, USA, **Bastián M. Fischer**, and **Sebastián Rodriguez**, Universidad Adolfo Ibáñez, Santiago, Chile

Objective: The present study examined the effectiveness of after-action reviews (AARs; also known as debriefing) in mitigating skill decay.

Background: Research on the long-term effectiveness of AARs is meager. To address this gap in the literature, we conducted an experimental study that also overcomes some research design issues that characterize the limited extant research.

Method: Eighty-four participants were randomly assigned to an AAR or non-AAR condition and trained to operate a PC-based fire emergency simulator. During the initial acquisition phase, individuals in the AAR condition were allowed to review their performance after each practice session, whereas individuals in the non-AAR condition completed a filler task. About 12 weeks later, participants returned to the lab to complete four additional practice sessions using a similar scenario (i.e., the retention and reacquisition phase).

Results: The performance of participants in the AAR condition degraded more after nonuse but also recovered faster than the performance of participants in the non-AAR condition, although these effects were fairly small and not statistically significant.

Conclusion: Consistent with the limited research on the long-term effectiveness of AARs, our findings failed to support their effectiveness as a decay-prevention intervention. Because the present study was conducted in a laboratory setting using a relatively small sample of undergraduate students, additional research is warranted.

Application: Based on the results of the present study, we suggest some additional strategies that trainers might consider to support long-term skill retention when using AARs.

Keywords: after-action reviews, debriefing, skill decay, mitigating loss, training

Address correspondence to Gonzalo J. Muñoz, Department of Psychology, Universidad Adolfo Ibáñez, Diagonal Las Torres 2640, Peñalolén, Santiago, RM, Chile; e-mail: gonzalo.munoz@uai.cl

HUMAN FACTORS
2022, Vol. 64(4) 760–778
DOI:10.1177/0018720820958848
Article reuse guidelines: sagepub.com/journals-permissions

An after-action review (AAR; also known as after-event review or debriefing) is a process whereby individuals or teams systematically review and discuss a recent performance event (e.g., Ellis & Davidi, 2005; Villado & Arthur, 2013). Several studies have confirmed the positive effects of AARs on performance across a variety of disciplines. Tannenbaum and Cerasoli's (2013) meta-analysis found that AARs resulted in a 25% performance improvement compared with control conditions, and Keiser and Arthur (2020) obtained even larger effects (d = 0.79). Similarly, Couper et al.'s (2013) review of the effectiveness of AARs in critical care settings found that AARs enhanced skill acquisition and transfer of skills related to managing life-threatening emergencies (e.g., resuscitation). For instance, 16 of 22 studies on technical performance outcomes (including clinical studies and simulations) were supportive of the use of AARs.

Whereas progress has been made toward understanding the design factors that influence the effectiveness of the AAR (e.g., review of successful versus failed performance [Ellis & Davidi, 2005], objective versus subjective reviews [Villado & Arthur, 2013], filmed versus personal event review [Ellis et al., 2010], colocated versus distributed reviews [Jarrett et al., 2016]), a limitation of this emerging body of literature is the paucity of studies that go beyond immediate post-training Keiser and Arthur (2020). Consequently, the goal of the present study was to extend the research on AARs by focusing on its long-term effectiveness. We accomplished this by using an experimental

design wherein the performances of participants in two conditions—AAR and non-AAR—were measured at several points, including a post-training performance assessment that took place approximately 12 weeks after the initial training.

THE AAR AS A DECAY-PREVENTION TRAINING INTERVENTION

Skill decay refers to observed decrements in acquired skills (or knowledge) after a period of nonuse (Arthur et al., 1998; Arthur & Day, 2020; Farr, 1987). Skill decay is particularly salient in situations where individuals do not regularly perform acquired skills or only perform these skills after extended periods of nonuse where they are expected to perform at full proficiency.

The distinction between long- and short-term memory is central to the phenomenon of decay. According to Bjork and Bjork (1992), long-term memory depends on storage strength or the extent to which memory representations are integrated with other memories, which facilitates the recovering of those memories after an extended period of nonuse. Because the extant research on memory posits that greater storage strength results from activities that entail processing information at a deeper level (Craik & Lockhart, 1972; Craik & Tulving, 1975), we advance that, to the extent that the AAR engenders psychological processes that promote deep processing, it should also enhance the retention of procedural and declarative knowledge that can help mitigate skill decay. Specifically, there are two components of the AAR that may reduce skill decay—self-explanation and data verification.

Self-explanation is "an active process of gathering, analyzing, and integrating data" (Ellis & Davidi, 2005, p. 858), which directs learners to reflect on their past behavior and facilitates the construction of "if-then" rules or mental models. During an AAR, participants are encouraged to generate self-explanations—to reflect on whether they met their goals, how their actions contributed to meeting those goals (or not), to set future objectives, and to reflect on strategies to improve on their past performance (Villado & Arthur, 2013). Because self-explanation involves processing information at a deeper level, we posit that this is an integral feature of the AAR that could reduce skill decay.

In addition to the deeper level of processing instigated by self-explanation, Chi et al. (1994) suggested that eliciting self-explanations also supports the development of more coherent mental models. Echoing this view, Ellis and Davidi (2005) posited that the AAR promotes deeper learning by encouraging participants to revise their mental models and integrate sometimes incompatible pieces of information into a coherent whole—a process that they labeled *data verification*. Consequently, yet another reason why the AAR might impact long-term retention is that it fosters the development of more coherent mental models of the task, which presumably facilitates retrieval (Willoughby & Wood, 1994).

PREVIOUS RESEARCH ON THE AAR AS A DECAY-PREVENTION INTERVENTION

Generally speaking, the long-term effectiveness of the AAR is indicated by the difference in post-training results between an experimental (AAR) and control condition after a specified period of time has elapsed. A direct comparison between post-training performance and pretraining performance may or may not be directly relevant, depending on the research context. For instance, post-training results may be higher than pretraining results if participants had further opportunities to practice the targeted skills.

Only a handful of studies have examined the long-term effectiveness of AARs (Levett-Jones & Lapkin, 2014). Morgan et al. (2009) conducted a randomized controlled trial to determine the effectiveness of AARs in the context of a high-fidelity simulation for training anesthetists to manage critical clinical cases. Morgan et al. found that AARs resulted in a modest but statistically significant 3.5% increase in posttest scores after 6–9 months of the initial training. In contrast, posttest scores of a control group remained unchanged after a similar nonuse period.

To the best of our knowledge, Morgan et al. (2009) is the only study that successfully isolated the effect of the AAR on post-training performance. However, due to the use of a single-practice session, it is possible that the small AAR effects observed in Morgan et al.'s study resulted from insufficient exposure to the training intervention, a validity threat that may lead to an incorrect characterization of the effectiveness of a treatment (Shadish et al., 2002). As a comparison, Wayne et al.'s (2005) study (on the effectiveness of AARs for training medical residents on advanced cardiac life support protocols) involved six practice sessions accompanied by the AAR. However, Wayne et al.'s study utilized a delayed-group design in which the control group did not have any hands-on practice with the simulator before the posttest session. Therefore, in Wayne et al.'s study, it is not possible to disentangle the effects of AARs from practice effects alone.

Other studies that focus on retention as a training outcome have been more concerned with investigating some variation of the AARs rather than on the effect of the AAR alone. For instance, using a 5-week nonuse period, Welke et al. (2009) compared the effectiveness of a personalized video-assisted oral AAR and a standardized computer-based multimedia AAR for training advanced cardiac life support among anesthesia residents. As another example, Chronister and Brown's (2012) study on critical care training for nursing students compared the effectiveness of a verbal AAR led by an expert to watching a video recording in addition to discussing their performance with the expert. However, because the comparison groups in Welke et al.'s (2009) and Chronister and Brown's (2012) studies received *some* form of the AAR, its long-term effectiveness per se cannot be inferred from these studies.

SUMMARY AND STUDY OVERVIEW

Based on our literature review of AARs, we contend that first, there is an overlap between the theoretical underpinnings of the AAR and the theories that have been posited to mitigate skill decay, which suggests that the effects of the AAR might extend beyond the initial acquisition period. Second, although scant, the empirical research on the effectiveness of AARs as a decay-prevention intervention tends to suggest that this intervention has a small effect on long-term performance, as shown in Morgan et al.'s (2009) study. However, the validity of these results is threatened by methodological and research design concerns that characterize these studies (i.e., lack of true control groups and insufficient exposure to the treatment). Hence, we contend that these deficiencies can be overcome by using a research design that has (1) an experimental group and a true control group; and (2) an experimental group that is exposed repeatedly to the AAR intervention. In addition, we posit that a sound methodological approach should consider measuring performance at several points during reacquisition to observe the effects of the AAR (or lack thereof) beyond the initial retention test. Despite its limitations, we posit that, as an initial step, the use of a laboratory setting allows us to determine more precisely whether or not the benefits of the AAR extend beyond initial training.

In the present study, participants were trained to operate *Fire Escape* (Muñoz et al., 2016), a PC-based synthetic task environment designed to train civilians on how to escape from a burning building. For brevity, we will refer to the initial acquisition phase as "Phase 1," and the retention and reacquisition phase as "Phase 2." Although the performance task was designed as an individual task, during Phase 1 participants were grouped into dyads for the purpose of conducting the AARs. However, when participants returned after the nonuse period, they completed the retention task alone. Phase 2 occurred about 12 weeks (M = 11.87 weeks, SD = 4.31) after Phase 1, a relatively long-time interval compared with the extant skill decay literature. (For instance, only three of the 111 samples of Wang et al., 2013 employed a nonuse interval of 90 days or greater. Therefore, in the context of academic research, we consider the present study to be a very stringent examination of the AAR as a skill-decay prevention intervention. However, we also acknowledge that in field settings, nonuse intervals of more than 12 weeks are common. For instance, laypersons may take part in emergency evacuation simulations a few times per year. If fire drills are conducted once a year,

the 12-week interval would seem short. Thus, the 12-week interval utilized here represents a compromise between the applied nature of the questions advanced and the potential to contribute to the scientific literature.) Because the focus of the study is on skill decay, we sought to determine (1) whether the amount of retention differs between individuals in the AAR and non-AAR conditions, wherein retention refers to the difference between scores from the last Phase 1 session to the first Phase 2 session; and (2) whether individuals in the AAR condition would reacquire lost skills faster than individuals in the non-AAR condition.

General Mental Ability, Psychomotor Ability, and Video-Game Experience as Control Variables

General mental ability (GMA), psychomotor skills, and video game experience were included as control variables. Because higher-ability individuals acquire more knowledge and skills than relatively lower-ability individuals, higher-ability individuals are also expected to maintain previously acquired knowledge and skills for longer periods of time (Day et al., 2013; Ree et al., 1995). Psychomotor ability was also included as a factor that may affect not only initial acquisition (e.g., Jarrett et al., 2017) but skill decay as well (for the same reason that GMA affects retention). Finally, because a person's general understanding of how video games work may reduce the cognitive load of learning a new game, we examined the extent to which video game experience affected acquisition performance and skill decay.

METHOD

Participants

The sample was comprised of college students from a variety of majors from a private university in Chile. This research complied with the tenets of the Declaration of Helsinki and was approved by the Institutional Review Board at University Adolfo Ibáñez. Informed consent was obtained from each participant. They were invited to participate in the study via e-mail and were offered course credit plus CLP$10 (approximately US$15) for their participation. The initial study sample consisted of 96 individuals (57% female) grouped into 48 dyads. However, only 84 participants (57% female) returned to the lab to complete Phase 2 and thus constituted the final study sample (mean age was 19.61 years, SD = 1.27). Assuming a medium effect size (f = .25) and a homogeneous correlation of .50 amongst the eight performance (repeated) measures, the power to detect a statistically significant effect for a sample of 84 participants was .86.

Measures

Performance task and scores. *Fire Escape* (Muñoz et al., 2016) was used as the performance task. *Fire Escape* is a cognitively complex, PC-based synthetic task environment that simulates a fire emergency in a high-rise building, and can be described as a combination of a first-person-shooter and complex-navigation game (Figure 1). The purpose of *Fire Escape* is to help trainees acquire knowledge and principles that are germane to evacuating a building on fire. The learning objectives of the simulation were developed in conjunction with two emergency experts with more than 15 years of experience as firefighters. The set of learning objectives (e.g., remain calm, call the fire department, find an escape route) was further validated by eight firefighters who rated the importance of the tasks using a five-point Likert scale (0 = *not relevant for civilians*; 1 = *somewhat important*; 2 = *moderately important*; 3 = *very important*; 4 = *extremely important*), and their difficulty to learn (0 = *not relevant for civilians*; 1 = *somewhat difficult to learn*; 2 = *moderately difficult to learn*; 3 = *very difficult to learn*; 4 = *extremely difficult to learn*). Mean importance ratings of the learning objectives ranged from 3.88 to 4.00, and difficulty to learn ranged from 2.63 to 3.00.

Participants operated *Fire Escape* using a monitor, keyboard, mouse, and headphones. Performance scores were obtained by adding the points of evacuation time (reverse-coded; *max* = 800 pts.), exiting the building (200 pts.), health points (*max* = 1000 pts.), door closing (10 pts. each), activating the alarm (100 pts.), extinguishing small fires (120 pts. each), and calling the fire department (100 pts.). Awarded points were based on the importance of the tasks. For instance, more points were awarded to maintaining health or

Figure 1. Fire Escape video game interface. The left-side panels show images of the game in the walkaround mode, whereas the right-side panels show the building after the start of the fire emergency. During the fire emergency, health and energy points are displayed using a red and green bar, respectively.

exiting the burning building rapidly than closing doors or activating the alarm.

Fire Escape sessions. During Phase 1, each participant completed a baseline performance session (Session 1) followed by four additional sessions (Sessions 2–5). Phase 2 consisted of four sessions (Sessions 6–9) but instead of the office theme used in Phase 1, the theme of Phase 2 was a hotel. The hotel scenario was designed to be equivalent in difficulty to the office scenario of Phase 1.

Before each session, participants were told that their objectives were to evacuate the building as quickly as possible, avoid the dangers they encounter along the way, and slow down the spread of fire without putting themselves at risk. After being informed of these objectives, they were allowed 1.5 min to walk around the building floor. After the 1.5 min had elapsed, participants had 5 min to complete each session. A session was terminated when either (1) participants successfully exited the building, (2) participants run out of health points (died), or (3) when the 5-min time limit expired. The simulator displayed the session runtime. Immediately after the session was over, performance scores appeared on the participant's screen.

Training Manipulation

After completing a test session, participants in the AAR condition participated in a 10-min AAR, facilitated by a trained research assistant. Prior to the first AAR, the facilitator explained the AAR process and provided a list of questions to be addressed during the AAR. During the AAR, the role of the facilitator was to ensure that participants had provided answers to each of the specified questions and that they completed the AAR within the 10-min time limit.

During the AAR, the research assistant asked the following questions to the participants: (1) What was the intended outcome? (2) What was the actual outcome? (3) What specific actions and behaviors contributed to meeting the intended outcome? (4) What specific actions and behaviors detracted from meeting the intended outcome? (5) What is the intended future outcome? (6) What actions will increase the likelihood of meeting the intended future

outcome? These questions were in accord with the stages of the AAR process advanced by Villado and Arthur (2013). Because each participant had completed the session individually, they took turns to provide answers to the specified questions. Importantly, we expected that responding to these questions would help trainees acquire the knowledge and understand the principles involved in the process of evacuating a building on fire. Furthermore, we expected that a systematic review of the intended outcomes (e.g., leave the building as fast as possible) and the strategies that could be employed to increase their scores during the simulation (e.g., avoid inhaling smoke to maintain health points) would encourage greater reflection and deeper understanding of the dynamics of the simulation, and thus increase retention of the acquired knowledge and principles.

Participants assigned to the non-AAR condition completed a filler task. The filler task consisted of reading literature passages (unrelated to the game) and answering questions to test their understanding.

Self-reported learning outcomes. As a manipulation check for the AAR, participants in both conditions completed the same six-item measure to assess the extent to which the goals of the AAR sessions were met (experimental condition) or how much they reflected on their experience using the simulator (control condition). A sample item from this measure was "I have a clear objective for the next session." The full set of items for this measure is presented in Table 3. Participants responded to each item using a five-point Likert scale (1 = *completely disagree*; 5 = *completely agree*). Cronbach's α for the resultant ratings were .69, .79, .76, and .83 for Sessions 1 to 4, respectively.

Video game experience. Video game experience was measured using the question "Overall, how would you rate your skill level for playing video games?" (c.f., Unsworth et al., 2015). Participants rated their skill level using a five-point scale (1 = *very little skill*; 5 = *very high skill*).

General mental ability. GMA was operationalized as scores on the letter sets (GMA-L; Ekstrom et al., 1976) and number series (GMA-N; Thurstone, 1938) tasks. Participants were given 7 min to complete 30 letter sets and 5 min to complete 15 number series. Scores were obtained by dividing the number of correct responses by the maximum possible score for each task (i.e., 30 and 15, respectively). Scores from these measures have been shown to correlate highly with scores from the Raven's Advanced Progressive Matrices (r = .63 with GMA-L; r = .63 with GMA-N) as well as the Wonderlic Personnel Test (r = .60 with GMA-L; r = .68 with GMA-N; see Hicks et al., 2015).

Psychomotor ability. A short psychomotor task that matched the elementary performance requirements of the simulator (i.e., pointing and clicking using a PC mouse) was developed by the first author and has been utilized in previous research (Arthur et al., 2015). A series of 4 × 4-cm colored targets appeared on different parts of a computer screen. The targets were either red or blue, and had the shape of a circle or a square. Participants were instructed to click as quickly as possible on either red circles or blue squares, but not on red squares or blue circles. After five practice trials, participants completed 20 consecutive trials. Scores were obtained by averaging participants' correct response times across the 20 trials. Thus, larger response times indicate lower psychomotor ability, and vice versa.

Procedure

Table 1 presents an overview of the research procedures. Participants were randomly assigned to the AAR or non-AAR conditions. At the beginning of Phase 1, participants completed a basic demographics questionnaire (e.g., sex, age, major), and the individual difference measures—video game experience, GMA, and psychomotor ability. Then, they began the first tutorial, which was designed to provide a basic understanding of how to operate the simulator.

Following the basic tutorial, participants completed their first session, which served as the baseline measure (Session 1). After Session 1, participants watched a 2-min video tutorial (i.e., the advanced tutorial) that showed several examples of how to interact

TABLE 1: Schedule of Activities for Each Session by AAR Condition

PHASE 1	
Schedule activities	
Informed consent	
Demographics	
Video game experience	
General mental ability	
Psychomotor ability	
Basic tutorial	
Session 1	
Advanced tutorial	
AAR	Non-AAR
Session 2	Session 2
AAR	Filler task
Self-reported learning outcomes	Self-reported learning outcomes
Session 3	Session 3
AAR	Filler task
Self-reported learning outcomes	Self-reported learning outcomes
Session 4	Session 4
AAR	Filler task
Self-reported learning outcomes	Self-reported learning outcomes
Session 5	Session 5
AAR	Filler task
Self-reported learning outcomes	Self-reported learning outcomes
PHASE 2 (about 12 weeks after Phase 1)	
Session 6	
Session 7	
Session 8	
Session 9	

Note. AAR = after-action review.

with the different objects they could encounter during the simulation (e.g., fire extinguisher, cellphone). Next, they completed Sessions 2–5. After each session, individuals in the experimental condition participated in the AAR, whereas individuals in the control condition completed a filler task. Phase 2 took place approximately 12 weeks after Phase 1 and consisted of four sessions. The first session of Phase 2 (Session 6) served as the focus of the retention analysis; subsequent scenarios (Sessions 7–9) were identical and served to model reacquisition.

RESULTS

Sample AAR Responses

In Table 2, we provide examples of participants' responses to the AAR questions. As a reminder, only participants in the AAR condition generated verbal responses; participants in the non-AAR condition only responded to the self-reported learning outcomes measure using a Likert scale.

Manipulation Checks

Whereas participants in *both* conditions had high scores on the learning-outcomes measure (above 4 on a 1–5-point scale), results from a *t* test for independent samples showed that the overall mean was higher for participants in the AAR condition following Session 2 ($t[82] = 4.04$, $d = 0.88$), Session 3 ($t[82] = 3.59$, $d = 0.78$), Session 4 ($t[82] = 4.04$, $d = 0.88$), and Session 5 ($t[82] = 4.71$, $d = 1.03$), $p < .05$ (two-tailed). Thus, as expected, the systematic appraisal of past performance promoted by the AAR resulted in noticeable differences in individuals' *perceptions* of their own learning. Mean differences between conditions for each individual item as well as for the overall measure are presented in Table 3.

Control Variables

To determine if the experimental (AAR) and control (non-AAR) conditions differed on the control variables, independent *t* tests were conducted, comparing the control variable means of the two conditions. The results indicated that the means of the experimental and control participants were not statistically significant for video game experience ($t[82] = 1.64$, $d = 0.36$), GMA-L ($t[82] = 0.79$, $d = 0.17$), GMA-N ($t[82] = 1.76$, $d = 0.38$), or psychomotor ability ($t[82] = 0.87$, $d = -0.19$), $p > .05$ (two-tailed). Consequently, they were not included as statistical controls in subsequent models.

TABLE 2: Sample Responses to the AAR Questions

AAR Questions	Sample Responses
(a) What was the intended outcome?	"Leave the building alive in the shortest possible time" "Leave with the highest health points" "Prevent the spread of fire"
(b) What was the actual outcome?	"I couldn't find the way out" "I didn't know how to use the fire extinguisher" "I went another way and got lost"
(c) What specific actions and behaviors contributed to meeting the intended outcome?	"I memorized the exits on the map" "I didn't waste time hanging around" "I ducked to not breathe the smoke"
(d) What specific actions and behaviors detracted from meeting the intended outcome?	"I wasted my time trying to use the fire extinguisher" "I wasted time trying to put out fires" "I tried to walk over the fire"
(e) What is the intended future outcome?	"Leave the building without losing health points" "Try harder to just leave the building, which is the main objective" "Improve my exit time"
(f) What actions will increase the likelihood of meeting the intended future outcome?	"Don't waste time trying things and try to get out as quickly as possible" "Be more careful, don't be so reckless" "Try to put out only the necessary fires"

Note. AAR = after-action review.

Main Analyses

This study utilized a 2 (training condition: AAR vs. non-AAR) × 9 (session: Sessions 1–9) repeated-measures design with training condition as the between-subjects independent variable and session as a within-subjects independent variable. Table 4 shows mean scores for all the study variables by condition (AAR and non-AAR) as well as their correlations. Although not statistically significant, there was a nontrivial performance difference between conditions at baseline (Session 1). To control for this difference and make the interpretation of results clearer, we used the baseline performance scores as a control. We accomplished this by centering Session 1 scores (i.e., subtracting each participant's score from the sample mean score) and then subtracting the centered scores from Sessions 2 to 9 performance scores. Figure 2 shows the performance scores by condition after controlling for Session 1 scores.

The following analyses are based on Bliese and Lang's (2016) recommendations for modeling discontinuous growth models, and were implemented via the *nlme* package in R (Version 3.1–148; Pinheiro et al., 2020). As shown in Table 5, a time variable was introduced to represent linear change during acquisition (TIME.A; where A stands for absolute). Note that TIME.A ranges from 0 to 3 during acquisition and is held constant at 3 during retention and reacquisition. As we explain later, the advantage of this particular time specification is that performance changes after the nonuse period can be interpreted in absolute terms. Two additional change variables were added to model the effect of the change event, which in the present study corresponds to the period of nonuse between the last session of Phase 1, and the first session of Phase 2. First, a transition variable (TRANS) was included to determine the degree to which the intercept was altered after the event. Here,

TABLE 3: Mean Ratings of Self-Reported Learning Outcomes Items After Sessions 2–5 by AAR and Non-AAR Conditions

	Session 2			Session 3			Session 4			Session 5		
Item	AAR *M* (*SD*)	Non-AAR *M* (*SD*)	*d*	AAR *M* (*SD*)	Non-AAR *M* (*SD*)	*d*	AAR *M* (*SD*)	Non-AAR *M* (*SD*)	*d*	AAR *M* (*SD*)	Non-AAR *M* (*SD*)	*d*
I understand what was the objective of the session	4.93 (0.25)	4.92 (0.27)	0.03	4.95 (0.21)	4.75 (0.49)	**0.55**	4.98 (0.15)	4.72 (0.72)	**0.50**	4.95 (0.21)	4.83 (0.45)	0.38
I clearly understand what was my result	4.84 (0.43)	4.5 (0.75)	**0.56**	4.84 (0.53)	4.60 (0.63)	0.42	4.93 (0.33)	4.72 (0.51)	**0.49**	4.95 (0.21)	4.58 (0.68)	**0.77**
I understand which actions contributed to achieving the objective	4.75 (0.49)	4.55 (0.75)	0.32	4.77 (0.60)	4.55 (0.71)	0.34	4.89 (0.39)	4.75 (0.49)	0.31	4.93 (0.33)	4.60 (0.55)	**0.74**
I understand what actions made it difficult to reach the objective	4.8 (0.41)	4.42 (0.84)	**0.57**	4.82 (0.45)	4.25 (1.03)	**0.73**	4.82 (0.39)	4.58 (0.68)	**0.45**	4.82 (0.58)	4.50 (0.75)	**0.48**
I have a clear objective for the next session	4.84 (0.43)	4.15 (0.89)	**1.00**	4.77 (0.48)	4.25 (0.98)	**0.69**	4.91 (0.29)	4.30 (0.99)	**0.85**	4.98 (0.15)	4.12 (0.91)	**1.34**
I know what actions to carry out to reach the future objective	4.48 (0.66)	4.00 (1.15)	**0.51**	4.61 (0.65)	4.17 (1.11)	**0.49**	4.77 (0.48)	4.17 (1.11)	**0.71**	4.77 (0.57)	4.20 (1.02)	**.70**
Overall	4.77 (0.26)	4.42 (0.50)	**0.88**	4.80 (0.31)	4.43 (0.59)	**0.78**	4.88 (0.20)	4.54 (0.52)	**0.88**	4.90 (0.26)	4.47 (0.54)	**1.03**

Note. $N = 84$ (AAR = 44; non-AAR = 40). A five-point rating scale (1 = *completely disagree*; 5 = *completely agree*) was used. *d* = standardized difference between self-reported learning outcomes by condition; AAR = after-action review. Values of *d* in bold are statistically significant ($p < .05$, two-tailed).

TABLE 4: Descriptives by Condition and Intercorrelations Between Study Variables

	Condition														
	AAR		Non-AAR												
	M	*SD*	*M*	*SD*	1	2	3	4	5	6	7	8	9	10	11
1. Condition	1.00	–	.00	–	–										
2. Nonuse interval (in days)	84.64	31.99	81.95	28.68	.04	–									
3. Sex	.45	.50	.40	.50	.06	−.20	–								
4. Age	19.64	1.46	19.57	1.03	.02	−.07	.35*	–							
5. Video game experience	3.56	1.16	3.12	1.24	.18	.02	.56*	.23*	–						
6. GMA-N	.60	0.15	.54	0.18	.19	−.04	.56*	.13	.40*	–					
7. GMA-L	.60	0.13	.58	0.12	.09	−.11	.24*	.02	.21	.43*	–				
8. Psychomotor ability	1165.40	157.77	1206.24	262.65	−.10	−.10	−.21	−.18	−.28*	−.23*	−.21	–			
9. LO S2	4.77	0.26	4.43	0.50	.41*	−.06	.18	.01	.20	.07	.06	−.09	–		
10. LO S3	4.80	0.31	4.43	0.59	.37*	−.01	.22*	.02	.28*	.36*	.24*	−.08	.61*	–	
11. LO S4	4.88	0.20	4.54	0.52	.41*	.01	.31*	.08	.29*	.32*	.08	−.22*	.52*	.73*	–
12. LO S5	4.90	0.26	4.47	0.54	.46*	.04	.00	−.02	.10	.19	.20	−.18	.29*	.48*	.59*
13. Performance S1	938.74	516.88	759.97	519.78	.17	.13	.28*	.12	.23*	.30*	.29*	−.20	.08	.11	.02
14. Performance S2	1103.70	567.81	966.67	522.77	.13	.09	.45*	.13	.31*	.28*	.17	−.25*	.22*	.29*	.30*
15. Performance S3	1263.86	512.33	1070.97	546.77	.18	.11	.34*	.01	.11	.29*	.21*	−.11	.21*	.27*	.27*
16. Performance S4	1363.86	500.92	1247.38	557.43	.11	.06	.48*	.09	.19	.27*	.20	−.14	.26*	.22*	.33*
17. Performance S5	1407.16	531.54	1292.90	540.70	.11	−.01	.48*	.12	.23*	.31*	.23*	−.12	.21*	.28*	.37*
18. Performance S6[c]	1300.68	512.07	1270.53	419.40	.03	−.11	.56*	.07	.25*	.35*	.21	−.06	.10	.10	.18
19. Performance S7	1455.70	419.46	1336.20	556.45	.12	−.09	.56*	.18	.45*	.28*	.27*	−.32*	.22*	.21	.39*
20. Performance S8	1526.91	414.18	1394.03	506.76	.14	−.18	.56*	.18	.39*	.32*	.33*	−.22*	.19	.15	.24*
21. Performance S9	1527.34	408.47	1439.56	501.44	.10	−.19	.52*	.16	.35*	.31*	.32*	−.30*	.17	.14	.24*

12	13	14	15	16	17	18	19	20	21

1. Condition
2. Nonuse Interval (in days)

(Continued)

TABLE 4 (Continued)

	12	13	14	15	16	17	18	19	20	21
3. Sex										
4. Age										
5. Video game experience										
6. GMA-N										
7. GMA-L										
8. Psychomotor ability										
9. LO S2										
10. LO S3										
11. LO S4										
12. LO S5	–									
13. Performance S1	−.08	–								
14. Performance S2	.15	.44*	–							
15. Performance S3	.11	.30*	.50*	–						
16. Performance S4	.13	.30*	.55*	.51*	–					
17. Performance S5	.38*	.10	.41*	.34*	.59*	–				
18. Performance S6[a]	.07	.25*	.41*	.30*	.32*	.38*	–			
19. Performance S7	.25*	.18	.44*	.35*	.44*	.43*	.46*	–		
20. Performance S8	.22*	.33*	.42*	.34*	.56*	.47*	.47*	.66*	–	
21. Performance S9	.15	.34*	.46*	.34*	.46*	.40*	.48*	.71*	.76*	–

Note. $N = 84$ (AAR = 44; non-AAR = 40). 0 = non-AAR; 1 = AAR. Dummy codes for sex are 0 = female and 1 = male; GMA measured via the number series task (*N*) or the letter sets task (L) *$p < .05$ (two-tailed). AAR = after-action review; LO = self-reported learning outcomes; GMA = general mental ability; S = session.

[a]S6 was the retention task.

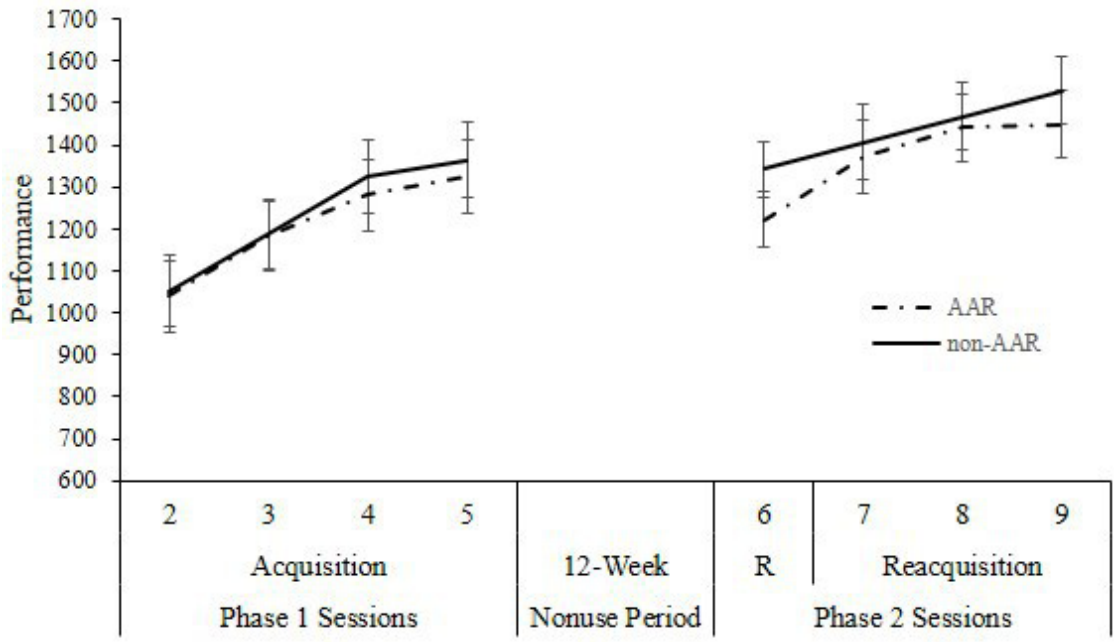

Figure 2. Mean performance and standard error of the mean by training condition (AAR vs. non-AAR) controlling for differences in baseline scores. $N = 84$ (AAR = 44; non-AAR = 40). AAR = after-action review; R = retention.

TABLE 5: Coding of Time Variables

Session	TIME.A	TRANS	RECOV
1	-	-	-
2	0	0	0
3	1	0	0
4	2	0	0
5	3	0	0
6	3	1	0
7	3	1	1
8	3	1	2
9	3	1	3

Note. Due to performance differences between conditions at baseline (Session 1), performance scores were transformed to control for said differences. Thus, the time variable (TIME.A) for Session 2 was set to 0. RECOV = recovery; TIME.A = time variable for examining absolute change; TRANS =transition.

time is coded 0 prior to the event and 1 after the event. Given that time during acquisition was coded using TIME.A, the TRANS variable is interpreted as the absolute change in performance after the nonuse period. Thus, if the coefficient for TRANS is −100, it means that performance decreased by 100 points after the nonuse period. Then, a recovery variable (RECOV) was added to determine the extent to which the postchange slope was different from 0. Phrased differently, the coefficient for RECOV is the rate of change during Phase 2.

Because the parameters thus obtained represent the prechange slope (TIME.A), transition (TRANS), and postchange slope (RECOV) for the entire sample, the last step for testing the study's hypotheses involves adding a dummy variable to represent each condition (0 = non-AAR group, 1 = AAR group). For interpretation purposes, the effectiveness of the AAR for reducing skill decay and facilitating skill reacquisition is indexed by the interaction terms between the training condition and the TRANS and RECOV variables, respectively.

As shown in Table 6, three models were estimated: a baseline model (Model 1); a second

TABLE 6: Longitudinal Random Coefficient Growth Models With Task Performance as the Dependent Variable

	Model 1		Model 2		Model 3	
Variable	*B*	*SE*	*B*	*SE*	*B*	*SE*
Intercept	1065.92*	65.98	1068.79*	96.82	772.66	516.78
Change predictors						
TIME.A	101.12*	19.74	106.63*	28.95	−407.36	231.98
TRANS	−69.62	49.51	−39.76	72.40	1506.88*	657.670
RECOV	66.73*	19.82	58.39*	29.13	−19.43	276.55
Predictors						
AAR			−5.45	132.86	−23.52	135.52
AAR × TIME.A			−10.35	39.70	−61.22	43.90
AAR × TRANS			−56.27	99.42	90.80	113.42
AAR × RECOV			15.67	39.87	8.81	45.90
Mediator						
LO					66.69	114.15
LO × TIME.A					114.57*	51.60
LO × TRANS					−345.24*	146.36
LO × RECOV					17.47	61.62
Variance components						
Intercept	229586.84		231949.57		217290.43	
Residual	149318.96		150312.55		148835.35	
−2 log likelihood	9626.85		9586.00		9525.72	
AIC	9640.85		9608.00		9555.72	
BIC	9672.09		9657.03		9622.48	
$R_t^{2(f_2)}$			.05		.44	
$R_b^{2(f_2)}$			.08		.57	

Note. N = 84 (AAR = 44; non-AAR = 40). TIME.A = dummy variable for indexing *absolute* change from baseline to end-of-acquisition performance; TRANS = dummy variable for the intercept change between the end-of-acquisition performance session (coded 0) and the first delayed performance session (coded 1); RECOV = dummy variable to test for postchange slope; 0 = non-AAR, 1 = AAR; $R_t^{2(f_2)}$ = Proportion of total outcome variance explained by level-2 predictors (AAR or LO) via fixed slopes; $R_b^{2(f_2)}$ = Proportion of between-individual outcome variance explained by level-2 predictors (AAR or LO) via fixed slopes (see Rights & Sterba, 2019); AAR = after-action review; AIC = Akaike's information criterion; BIC = Bayesian information criterion; LO = self-reported learning outcomes.

model (Model 2) for testing the effect of the AAR on the rate of acquisition, decay, and recovery; and a third model (Model 3) to examine the role of self-reported learning outcomes (this model is introduced in the "Supplementary analyses" section). First, we estimated a model (Model 1) using only the change parameters (TIME.A, TRANS, and RECOV). As indicated by the

likelihood ratio (*LR*) test, including a term to account for autocorrelation improved model fit significantly, $LR = 21.12$, $p < .05$. However, evidence of heteroscedasticity after accounting for autocorrelation was not found. Finally, including the random effects for TIME.A, TRANS, and RECOV did not improve model fit, which indicates no substantial differences between individuals in their rate of acquisition, decay, or recovery. Consequently, Model 1 as well as subsequent models were estimated accounting for autocorrelation only (see Bliese & Ployhart, 2002).

As shown in Table 6, results from Model 1 indicate that participants' scores increased during the acquisition stage, decreased following the nonuse period, and increased again during reacquisition. However, only the coefficients for TIME.A and RECOV were statistically significant. Thus, in absolute terms, the observed drop in performance following the nonuse period was not statistically significantly different from 0.

In Model 2, we evaluated the extent to which the condition effect interacted with the TRANS and RECOV variables for predicting performance. Table 6 shows that none of the interaction terms were statistically significant. Thus, the AAR was ineffective at reducing skill decay and did not facilitate the reacquisition of previously acquired skills. However, it is informative to interpret the observed parameters for descriptive purposes. Results from Model 2 show that the AAR × TRANS parameter was –56.27, which suggests that participants' scores in the AAR condition degraded faster than those of participants in the non-AAR condition. Also, the AAR × RECOV parameter ($B = 15.67$) indicates that participants in the AAR condition recovered faster than participants in the non-AAR. As can be seen in the lower part of Table 6, the proportion variance explained relative to either the total (.05) or inter-individual variance (.08) was fairly low. However, we reiterate that statistical significance tests do not support the proposition that the AAR mitigated skill decay or that it facilitated the reacquisition of previously acquired skills.

Supplementary analyses. As previously mentioned, the self-reported learning outcomes measure scores were higher for the AAR participants compared with the non-AAR participants. Prompted by one of the reviewers, we explored the role of self-reported learning outcomes as a mediator between the AAR and performance during acquisition, retention, and reacquisition. First, we tested a model using the self-reported learning outcomes measure as an outcome in a model with two predictors, namely, TIME.A and the dummy variable used to represent each condition (0 = non-AAR group, 1 = AAR group). Results of this step showed that individuals in the AAR condition had higher self-reported learning outcomes than individuals in the non-AAR condition, $B = 0.38$, $SE = 0.07$, $t(82) = 5.63$, $p < .05$. Then, we tested a model using self-reported learning outcomes and its interaction with the time variables. Given that the self-reported learning outcomes measure was obtained during acquisition (Sessions 2–5), we used the mean of this measure as a predictor for the retention and reacquisition sessions (Sessions 6–9). Results from Model 3 in Table 6 indicate that the interaction between self-reported learning outcomes and TIME.A was statistically significant, $B = 114.57$, $SE = 51.60$, $t(554) = 2.22$, $p < .05$. In addition, according to Model 3, there was a statistically significant interaction between self-reported learning outcomes and TRANS, $B = -345.24$, $SE = 146.36$, $t(554) = -2.36$, $p < .05$. Together, these results indicate that individuals with higher self-reported learning outcomes increased their performance at a higher rate during acquisition but also showed a more abrupt decline in performance following the nonuse period. Importantly, the proportion of variance explained by the self-reported learning outcomes measure was fairly high (.44 and .57 relative to the total and the inter-individual variance, respectively).

Figure 3 shows the mean performance of individuals with relatively higher and lower self-reported learning outcomes scores. Consistent with the results of Model 3, as shown in Figure 3, individuals above the median on the self-reported learning outcomes measure obtained higher scores during acquisition but their performance after the nonuse period decreased markedly. Thus, those who appeared to learn the most during acquisition were also the ones who had *more* to lose after the nonuse period.

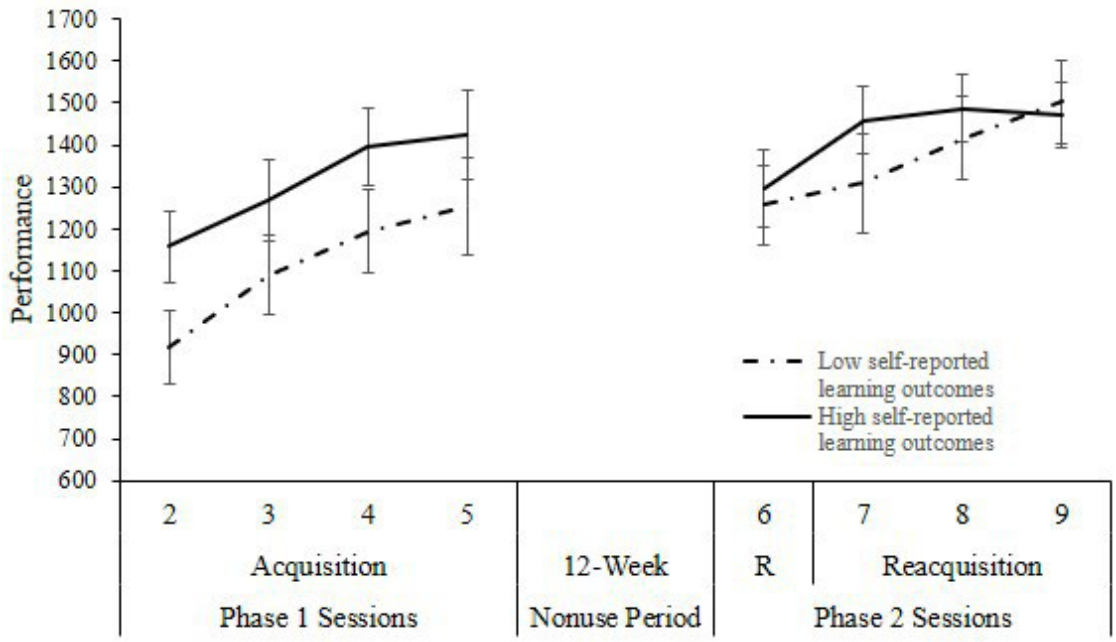

Figure 3. Mean performance of individuals above and below the median on the self-reported learning outcomes measure during acquisition, retention, and reacquisition. $N = 84$ (low self-reported learning outcomes = 41; high self-reported learning outcomes = 43). R = retention.

DISCUSSION

Consonant with the extant, albeit limited, research on the effectiveness of AARs as a decay-prevention intervention, the results from the present study showed that AARs had no effect on retention. In fact, the performance of individuals in the AAR condition decayed somewhat faster than the performance of the control group. At the same time, individuals in the AAR condition demonstrated a more rapid—although not statistically significant—recovery than individuals in the non-AAR group. Subsequent analyses suggested that the effect of the AAR on performance was mediated by self-reported learning outcomes. Specifically, the present findings indicate that the increase in performance during acquisition depends on the extent to which individuals learned from the AAR. Yet, results also showed that the performance of individuals who benefited the most from the AAR decayed faster after the nonuse period.

The effectiveness of emergency training depends on individuals acquiring knowledge and principles that must be remembered for long periods of time. Although the emergency simulator involves executing (virtual) behaviors on a PC, it is expected that carrying out the simulation tasks will facilitate the acquisition of procedural knowledge specific to the emergency training context. The point is that acquiring procedural knowledge is a cognitive activity. This is important for two reasons. First, cognitive skills are more prone to skill decay (e.g., Arthur et al., 1998). Thus, if such cognitive skills are the focus of laypersons' emergency training, then it is certainly relevant to design training to prevent skill decay. Second, by promoting deep processing, the AAR should be well-suited to support cognitive tasks. However, based on the results of the present study, it is unclear whether its benefits extend beyond the initial acquisition period.

Limitations and Future Directions

It needs to be acknowledged that a serious drawback of the retention literature is the limited number of empirical studies based on

samples of experts (Arthur & Day, 2020). For the purposes of the present study, the use of a nonexpert sample (i.e., college students) is not inappropriate because the performance task was designed to train civilians and the said civilians would, of course, be nonexperts. Although it would be imprudent to generalize the findings of this particular study to samples of experts, it is important to note that the same results have been obtained with samples of true experts. For instance, Morgan et al. (2009) used a sample of certified anesthetists and found that AARs resulted in a modest effect on retention after a long nonuse period. Thus, the limited research there is on this issue suggests that the AAR does not affect performance beyond the initial acquisition period, either using experts or nonexperts.

Despite the present study's relatively small sample size, it had sufficient power to detect a moderate effect between conditions (see "Participants" section) and, by extension, to detect effects other than the main one of interest—such as differences between conditions on the self-reported learning outcomes measure, and associations between individual difference variables and performance scores. Although there is no objective criterion for judging the appropriateness of a particular experiment before accepting a null finding (Frick, 1995), we posit that the methodology of the present study at least increased the likelihood of finding an effect (e.g., many trials per participant, using a laboratory setting, increasing the strength of the manipulation). Furthermore, the small, nonstatistically significant effects observed in the present study's results are consistent with previous research (e.g., Welke et al., 2009). Thus, instead of focusing on whether or not the AAR affects retention, it seems that the focus of future research should be how to ensure that the benefits of the AAR extend beyond the acquisition stage.

Whereas we acknowledge that the present findings may be limited to the specific manner in which the AAR was delivered, alternative frameworks (e.g., Tannenbaum & Cerasoli, 2013) have substantial overlap with the one utilized in the present study. Nevertheless, we posit that some key training design features of the AAR (e.g., level of involvement and characteristics of the instructor, AAR timing, and duration) have yet to be informed by research.

The self-reported learning outcomes measure was administered to participants in the control condition multiple times, which may have worked unintentionally as a lessened version of the AAR. Although the psychological processes engendered by the AAR are notoriously different from providing a quick answer on a five-point scale, the magnitude of the effect of administering the self-reported learning outcomes measure could be further investigated using a Solomon four-group design. In this design, half the participants in each condition (AAR and non-AAR) would complete the measure whereas the other half would not. Then, performance differences between these groups would allow researchers to directly determine the size of this effect.

In contrast to participants in the AAR condition, participants in the control group completed a filler task between sessions—that is, reading literature passages and answering questions to test their understanding—to maintain participants' focus off-task. The reason for including this particular task was to ensure that at the time of the last session, the participants in both conditions had exerted the same cognitive effort. Otherwise, by the end of the acquisition phase, participants in the AAR would have spent 40 additional on-task minutes than participants in the control group, which may have limited their cognitive resources to acquire new knowledge at later stages. As pointed out by one of the reviewers, in previous studies (e.g., Villado & Arthur, 2013), participants in the control group go from one session to the next virtually without pause, which drastically limits their chances to reflect on past performance in any meaningful way. Thus, yet another limitation of the present study design is that participants in the control condition may have been reflecting on their performance, which may account for the lack of difference between experimental conditions.

One significant advantage of AARs is that they can be easily embedded into initial training, and hence they do not engender either the logistic challenges or the costs associated with post-training interventions in which participants

need to be either fully retrained (i.e., hands-on training) or avail themselves to a lessened version of the initial training (observation rehearsal; Villado et al., 2013) during the nonuse period. However, it appears that AARs do not inherently promote the "transfer appropriate" processing (e.g., retrieval practice) that supports long-term learning. For instance, although the AAR may promote more active engagement in understanding the task at hand, if individuals do not systematically try to remember the lessons learned from the AAR (i.e., retrieval practice), then whatever they learn will likely be forgotten. A recommendation that arises from the present study is that the AAR should probably be supplemented with a generative learning strategy that supports long-term learning, such as asking participants to summarize the lessons learned during the AAR at the end of each session or asking them to retrieve those lessons during the nonuse period.

ACKNOWLEDGMENTS

We gratefully thank Álvaro Mardones (Emergency Physician and Senior Firefighter) and Cristóbal Mena (Deputy Director for the National Emergency Office, Ministry of the Interior and Security, Chile) for their expert advice during the development of the emergency simulator used in the present study. We also thank Jorge Villalón from the School of Engineering and Science of Universidad Adolfo Ibáñez for his continuous guidance and support to the programming team. This work was supported by FONDECYT Chile under Grant No 11140488.

KEY POINTS

- Whereas theory and research support the use of AARs as a training intervention, a noticeable gap in the literature is the paucity of studies examining the long-term effectiveness of AARs. The present study begins to address this gap.
- In contrast to previous research that relies on a single performance test during initial acquisition and a single performance test during delayed posttraining, in the present study participants' performance was assessed at several points before and after the nonuse period, which permitted a more comprehensive examination of the effectiveness of AARs in reducing skill decay and facilitating skill reacquisition.
- Findings from the present study showed that the AAR did not reduce skill decay. Furthermore, trainees who appeared to benefit more from the AAR during initial acquisition were also the ones whose performance suffered the most after the nonuse period. We suggest that additional generative learning strategies may be needed to reap the benefits of the AAR in the long term.

REFERENCES

Arthur, W., Jr., Bennett, W., Jr., Stanush, P. L., & McNelly, T. L. (1998). Factors that influence skill decay and retention: A quantitative review and analysis. *Human performance*, *11*, 57–101. https://doi.org/10.1207/s15327043hup1101_3

Arthur, W., Jr., & Day, E. A. (2020). Skill decay: The science and practice of mitigating skill loss and enhancing retention. In P. Ward, J. M. Schraagen, J. Gore, & E. Roth (Eds.), *The oxford handbook of expertise: Research & application* (pp. 1085–1108). Oxford University Press.

Arthur, W., Jr., Naber, A. N., Muñoz, G. J., McDonald, J. N., Atoba, O. A., Cho, I., Keiser, N. L., C. D, W., Glaze, R. M., Jarrett, S. M., Schurig, I., & Bennett, W., Jr. (2015). *An investigation of skill decay and reacquisition of individual- and team-based skills in a synthetic training environment* [Conference session]. American Psychological Association Division 19 Suite presentation at the 123rd Annual Convention of the American Psychological Association, Toronto, Ontario, Canada.

Bjork, R. A., & Bjork, E. L. (1992). A new theory of disuse and an old theory of stimulus fluctuation. In A. Healy, S. Kosslyn, & R. Shiffrin (Eds.), *From learning processes to cognitive processes: Essays in honor of William K. Estes* (Vol. 2, pp. 35–67). Erlbaum.

Bliese, P. D., & Lang, J. W. B. (2016). Understanding relative and absolute change in discontinuous growth models: Coding alternatives and implications for hypothesis testing. *Organizational Research Methods*, *19*, 562–592.

Bliese, P. D., & Ployhart, R. E. (2002). Growth modeling using random coefficient models: Model building, testing, and illustrations. *Organizational Research Methods*, *5*, 362–387. https://doi.org/10.1177/109442802237116

Chi, M. T. H., de Leeuw, N., Chiu, M. -H., & LaVancher, C. (1994). Eliciting self-explanations improves understanding. *Cognitive Science*, *18*, 439–477.

Chronister, C., & Brown, D. (2012). Comparison of simulation Debriefing methods. *Clinical Simulation in Nursing*, *8*, e281–e288. https://doi.org/10.1016/j.ecns.2010.12.005

Couper, K., Salman, B., Soar, J., Finn, J., & Perkins, G. D. (2013). Debriefing to improve outcomes from critical illness: A systematic review and meta-analysis. *Intensive Care Medicine*, *39*, 1513–1523. https://doi.org/10.1007/s00134-013-2951-7

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

Craik, F. I. M., & Tulving, E. (1975). Depth of processing and the retention of words in episodic memory. *Journal of Experimental Psychology: General*, *104*, 268–294. https://doi.org/10.1037/0096-3445.104.3.268

Day, E. A., Arthur, W., Jr., Villado, A. J., Boatman, P. R., Kowollik, V., Bhupatkar, A., & Bennett, W., Jr. (2013). Relating individual differences in ability, personality, and motivation to the retention and transfer of skill on a complex command-and-control simulation task. In W. Arthur, Jr., E. A. Day,

W. Bennett, Jr., & A. M. Portrey (Eds.), *Individual and team skill decay: The science and implications for practice* (pp. 282–301). Routledge.
Ekstrom, R. B., French, J. W., Harman, M. H., & Demen, D. (1976). *Manual kit of factor referenced cognitive tests*. Educational Testing Service.
Ellis, S., & Davidi, I. (2005). After-event reviews: Drawing lessons from successful and failed experience. *Journal of Applied Psychology*, *90*, 857–871. https://doi.org/10.1037/0021-9010.90.5.857
Ellis, S., Ganzach, Y., Castle, E., & Sekely, G. (2010). The effect of filmed versus personal after-event reviews on task performance: The mediating and moderating role of self-efficacy. *Journal of Applied Psychology*, *95*, 122–131. https://doi.org/10.1037/a0017867
Farr, M. J. (1987). *The long-term retention of knowledge and skill: A cognitive and instructional perspective*. Springer-Verlag.
Frick, R. W. (1995). Accepting the null hypothesis. *Memory & Cognition*, *23*, 132–138.
Hicks, K. L., Harrison, T. L., & Engle, R. W. (2015). Wonderlic, working memory capacity, and fluid intelligence. *Intelligence*, *50*, 186–195. https://doi.org/10.1016/j.intell.2015.03.005
Jarrett, S. M., Glaze, R. M., Schurig, I., & Arthur, W., Jr. (2017). The importance of team sex composition in Team-Training research employing complex psychomotor tasks. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *59*, 833–843. https://doi.org/10.1177/0018720816689744
Jarrett, S. M., Glaze, R. M., Schurig, I., Muñoz, G. J., Naber, A. M., McDonald, J. N., Bennett, W., Jr., & Arthur, W., Jr. (2016). The comparative effectiveness of distributed and colocated team after-action reviews. *Human Performance*, *29*, 408–427. https://doi.org/10.1080/08959285.2016.1208662
Keiser, N. L., & Arthur, W., Jr.. (2020). A meta-analysis of the effectiveness of the after-action review (or debrief) and factors that influence its effectiveness. *The Journal of applied psychology*, *Advance online publication*. https://doi.org/10.1037/apl0000821
Levett-Jones, T., & Lapkin, S. (2014). A systematic review of the effectiveness of simulation debriefing in health professional education. *Nurse Education Today*, *34*, e58–e63. https://doi.org/10.1016/j.nedt.2013.09.020
Morgan, P. J., Tarshis, J., LeBlanc, V., Cleave-Hogg, D., DeSousa, S., Haley, M. F., Herold-McIlroy, J., & Law, J. A. (2009). Efficacy of high-fidelity simulation debriefing on the performance of practicing anaesthetists in simulated scenarios. *British Journal of Anaesthesia*, *103*, 531–537. https://doi.org/10.1093/bja/aep222
Muñoz, G. J., Cortéz, D. A., Álvarez, C. B., Raggio, J. A., Concha, A., Rojas, F., Fischer, B. M., & Rodriguez, S. (2016). *Fire Escape [Computer software]*.
Pinheiro, J., Bates, D., DebRoy, S., & Sarkar, D., & R Core Team. (2020). *nlme: linear and nonlinear mixed effects models* (R package version 3.1-148). Retrieved from https://CRAN.R-project.org/package=nlme
Ree, M. J., Carretta, T. R., & Teachout, M. S. (1995). Role of ability and prior knowledge in complex training performance. *Journal of Applied Psychology*, *80*, 721–730. https://doi.org/10.1037/0021-9010.80.6.721
Rights, J. D., & Sterba, S. K. (2019). Quantifying explained variance in multilevel models: An integrative framework for defining R-squared measures. *Psychological Methods*, *24*, 309–338. https://doi.org/10.1037/met0000184
Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Wadsworth Cengage Learning.
Tannenbaum, S. I., & Cerasoli, C. P. (2013). Do team and individual debriefs enhance performance? A meta-analysis. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *55*, 231–245. https://doi.org/10.1177/0018720812448394
Thurstone, L. L. (1938). *Primary mental abilities*. University of Chicago Press.
Unsworth, N., Redick, T. S., McMillan, B. D., Hambrick, D. Z., Kane, M. J., & Engle, R. W. (2015). Is playing video games related to cognitive abilities? *Psychological Science*, *26*, 759–774. https://doi.org/10.1177/0956797615570367
Villado, A. J., & Arthur, W., Jr. (2013). The comparative effect of subjective and objective after-action reviews on team performance on a complex task. *Journal of Applied Psychology*, *98*, 514–528. https://doi.org/10.1037/a0031510
Villado, A. J., Day, E. A., Arthur, W., Boatman, P. R., Kowollik, V., Bhupatkar, A., & Bennett, W., Jr. (2013). Use of, reaction to, and efficacy of observational rehearsal training: Enhancing skill retention on a complex command-and-control simulation. In W. Arthur, Jr., E. A. Day, W. Bennett, Jr., & A. M. Portrey (Eds.), *Individual and team skill decay: The science and implications for practice* (pp. 240–257). Routledge.
Wang, X., Day, E. A., Kowollik, V., Schuelke, M. J., & Hughes, M. G. (2013). Factors influencing knowledge and skill decay after training. In W. Arthur, Jr., E. A. Day, W. Bennett, Jr., & A. M. Portrey (Eds.), *Individual and team skill decay: The science and implications for practice* (pp. 68–116). Routledge.
Wayne, D. B., Butter, J., Siddall, V. J., Fudala, M. J., Linquist, L. A., Feinglass, J., Wade, L. D., & McGaghie, W. C. (2005). Simulation-based training of internal medicine residents in advanced cardiac life support protocols: A randomized trial. *Teaching and Learning in Medicine*, *17*, 202–208. https://doi.org/10.1207/s15328015tlm1703_3
Welke, T. M., LeBlanc, V. R., Savoldelli, G. L., Joo, H. S., Chandra, D. B., Crabtree, N. A., & Naik, V. N. (2009). Personalized oral debriefing versus standardized multimedia instruction after patient crisis simulation. *Anesthesia & Analgesia*, *109*, 183–189. https://doi.org/10.1213/ane.0b013e3181a324ab
Willoughby, T., & Wood, E. (1994). Elaborative interrogation examined at encoding and retrieval. *Learning and Instruction*, *4*, 139–149. https://doi.org/10.1016/0959-4752(94)90008-6

Gonzalo J. Muñoz is an assistant professor in the Department of Psychology at University Adolfo Ibáñez, Chile. His PhD is in industrial-organizational psychology from Texas A&M University.

Diego A. Cortéz is a research assistant at Training Lab UAI. He received a bachelor's degree in psychology from Universidad Adolfo Ibáñez.

Constanza B. Álvarez is a research assistant at Training Lab UAI. She received her master's degree in organizational psychology from Universidad Adolfo Ibáñez.

Juan A. Raggio is a research assistant at Training Lab UAI. He received his master's degree in organizational psychology at Universidad Adolfo Ibáñez.

Antonia Concha is a research assistant at Training Lab UAI. She received her master's degree in organizational psychology from Universidad Adolfo Ibáñez.

Francisca I. Rojas is a research assistant at Training Lab UAI. She received her master's degree in organizational psychology from Universidad Adolfo Ibáñez.

Winfred Arthur, Jr. is a professor of psychology and management at Texas A&M University. His PhD is in industrial-organizational psychology from the University of Akron.

Bastián M. Fischer is a research assistant at Training Lab UAI. He received his MSc in engineering in 2018,

with an emphasis in information technologies, from Universidad Adolfo Ibáñez.

Sebastián Rodríguez is a research assistant at Training Lab UAI. He received his MSc in engineering, with an emphasis in information technologies, from Universidad Adolfo Ibáñez.

Date received: December 18, 2018
Date accepted: August 24, 2020