DOT/FAA/TC-22/11

TNO 2022 R10114

Federal Aviation Administration
William J. Hughes Technical Center
Aviation Research Division
Atlantic City International Airport
New Jersey 08405

Negative transfer of training of suboptimal degrees of variability in the training of procedures

NOTICE

This document is disseminated under the sponsorship of the U.S. Department of Transportation in the interest of information exchange. The U.S. Government assumes no liability for the contents or use thereof. The U.S. Government does not endorse products or manufacturers. Trade or manufacturers' names appear herein solely because they are considered essential to the objective of this report. The findings and conclusions in this report are those of the author(s) and do not necessarily represent the views of the funding agency. This document does not constitute FAA policy. Consult the FAA sponsoring organization listed on the Technical Documentation page as to its use.

This report was originally published by the Netherlands Organization for Applied Scientific Research (TNO), Human Performance; Soesterberg, The Netherlands as TNO 2022 R10114. The FAA has not edited or modified any content of this report.

This report is available at the Federal Aviation Administration William J. Hughes Technical Center's Full-Text Technical Reports page: actlibrary.tc.faa.gov in Adobe Acrobat portable document format (PDF).

Technical Report Documentation Page

Form DOT F 1700.7 (8-72) Reproduction of completed page authorized

1. Report No.	2. Government Accession No.	3. Recipient's Catalog No.
DOT/FAA/TC-22/11		

4. Title and Subtitle	5. Report Date
Negative transfer of training of suboptimal degrees of variability in the training of procedures	May 2022
	6. Performing Organization Code

7. Author(s)	8. Performing Organization Report No.
A. Landman, H. Pennings, R. Blankendaal, K. van den Bosch and E. Groen	

9. Performing Organization Name and Address	10. Work Unit No. (TRAIS)
Netherlands Organization for Applied Scientific Research (TNO), Human Performance; Soesterberg, The Netherlands.	TNO 2022 R10114
	11. Contract or Grant No.

12. Sponsoring Agency Name and Address	13. Type of Report and Period Covered
FAA Policy and Innovation Division Aircraft Certification Service Attn: Jeff Schroeder Bldg. N243 Moffett Field, CA 94035	
	14. Sponsoring Agency Code

15. Supplementary Notes

This report was originally published by the Netherlands Organization for Applied Scientific Research (TNO), Human Performance; Soesterberg, The Netherlands as TNO 2022 R10114

16. Abstract

Many professions require procedural skills to perform routine tasks, as well as adaptive skills to deal with unexpected situations. Whereas routine skills are acquired by repetitive practice of the same task, adaptive skills develop with variability of practice. We hypothesize that negative transfer may occur when routine skills are trained with (too) much variability or when adaptive skills are trained with (too) little variability.

The current study investigates the effect of different degrees of variability in the training of procedures on the transfer of training. Using the serious game "Space Fortress", 76 paid volunteers were trained how to manually control a spacecraft, while destroying as many hostile space fortresses as possible by firing shots. For the purpose of this study, we added another task, which involved the dismantling of "mines" by means of a multi-step procedure, i.e., pressing a specific series of keys.

All participants received the same Basic skills training (controlling the game). The Procedure training (dismantling mines) was varied between three groups. A "Low-Var" group: practiced one dismantling procedure per training session. A "Med-Var" group: practiced separate procedures in the initial part of training, followed by more mixed practice sets as training progressed. A "High-Var" group practiced the procedures in mixed order right from the onset of training.

We found that that the Low-Var group needed the most time to dismantle the mines on a post-training test with mixed procedures. They also made more errors on initiating the procedure. The High-Var group however made the most errors on a new mine test (requiring a procedure not trained before) performing significantly worse than the Low-Var group. These effects appeared only directly after the training, but were no longer present in a retention test given one week later. These results indicate that "too much" as well as "too little" variability of practice can lead to negative transfer of training depending on the context in which the procedures are performed.

17. Key Words	18. Distribution Statement
Training procedures, variability in training, training transfer, training retention, negative training	This document is available to the U.S. public through the National Technical Information Service (NTIS), Springfield, Virginia 22161. This document is also available from the Federal Aviation Administration William J. Hughes Technical Center at actlibrary.tc.faa.gov.

19. Security Classif. (of this report)	20. Security Classif. (of this page)	21. No. of Pages	19. Security Classif. (of this report)
Unclassified	Unclassified	39	Unclassified

Contents

Figures

Tables

Executive summary

Many professions require good procedural skills to perform routine tasks, as well as adaptive skills to deal with new or unexpected situations. It is known that routine skills are primarily developed by repetitively practicing the same task, whereas adaptive skills are developed by variability of practice. However, it is unknown how much variability should be included in training. It is hypothesized that "too much" variability will lead to negative transfer to situations requiring routine skills. Similarly, "too little" variability in the training can cause negative transfer in situations requiring adaptive skills. This study investigates the effect of different degrees of variability during the practice of procedures on the transfer of training.

The experiment used a web-based interface to administer the training and tests. The serious game "Space Fortress" was used as task to be learned. The primary task in the game is to control a spacecraft, and destroy as many hostile space fortresses as possible by firing shots. For the purpose of this study, we added another task, involving the dismantling of "mines" by means of a multi-step procedure, i.e., pressing a specific series of keys. The participants practiced dismantling three different types of mines, each requiring a different procedure. The experiment consisted of two training blocks of about two hours each (Basic skills training, and Procedure training), and two tests of about 30 minutes each (a Direct test immediately after completion of the training, and a Retention test one week later). Both tests consisted of a Repetitive test (one procedure), a Variable test (different procedures), and a New mine test (new, untrained procedure, consisting of a combination of two trained procedures).

In total, the training and test sessions were completed by 76 paid volunteers (56.6% male), aged between 18 and 45 years (M=26.6). All participants received the same Basic skills training (controlling the game). The Procedure training (dismantling mines) was varied between three groups. A "Low-Var" group: practiced one dismantling procedure per training session. A "Med-Var" group: practiced separate procedures in the initial part of training, followed by more mixed practice sets as training progressed. A "High-Var" group practiced the procedures in mixed order right from the onset of training.

On a post-training test with mixed procedures, the Low-Var group needed significantly more time than the other groups to dismantle the mines, and made more errors on initiating the procedure. On the New mine test (requiring a procedure not trained before), the High-Var group made the most errors, and performed significantly worse than the Low-Var group. These effects appeared only directly after the training, but were no longer present in a retention test, one week later.

All groups performed equal on a repetitive test that required execution of the separate procedures in a fixed order. In addition, all groups performed equally well on the post-training test on the basic skills.

We found evidence of negative transfer of training for low- as well as high variability of practice, both affecting performance on the procedure task, but in a different context. Low variability in procedure training led to suboptimal performance in *selecting* the correct procedure. High variability in procedure training led to suboptimal performance in executing untrained procedures.

The results are relevant for designing the training of pilots, where training time is usually very restricted, and instructors must decide how to balance repetitive and variable training.

1 Introduction

Negative transfer of training refers to situations in which learning in a formal training environment results in a degradation of performance in a new, or operational environment (Alexander, Brunyé, Sidman, & Weil, 2005) (Borgvall & Nählinder, 2008) (Burke, 1997) (Woltz, Gardner, & Bell, 2000). This is a problem in particular for the training of professionals, e.g., in the aviation or medical domain, where errors in task execution may have serious consequences. As an example of negative transfer, one can think of the transition to a new aircraft type, where pilots may need more time to adapt to new procedures, which are slightly different from the procedures they had practiced in their previous aircraft type.

As explained in our previous report on negative transfer of training (Pennings, Oprins, Schoevers, & Groen , 2019), training experts often associate negative transfer of training with difficulties in finding a balance between becoming a "routine expert" (ability to handle normal procedures, psychomotor skills) versus an "adaptive expert" (ability to also deal with non-normal situations). Variability of practice in training is important for the development of adaptive skills (Ford & Schmidt, 2000) (Landman, et al., 2018) (Van Merriënboer & Kirschner, 2007), although the acquisition of adaptive skills also benefits from repetitive practice to some extent (Arthur, Bennett Jr, Stanush, & McNelly, 1998) (Van Merriënboer, Kester, & Paas, 2006). The challenge is thus to find a balance between repetition and variability in training, especially when training time is limited, which usually is the case in professional training. It can be expected that disturbing the balance to either side ("too little" or "too much") will have negative consequences for the training results. Similarly, inappropriate timing (e.g., "too early") of variability may also hamper the acquisition of skills.

This study investigates the effectiveness of training with different degrees of repetition and variability. It is hypothesized that a high degree of variability (i.e., high contextual interference) will negatively affect the trainees' routine skills, whereas a low degree of variability (i.e., low contextual interference) will hamper the acquisition of adaptive skills. These hypotheses are tested by comparing the performance of three participant groups, who receive different degrees of repetition and variability in their training. It can be argued that this approach does not precisely test "negative transfer of training", as defined in the first paragraph of this introduction. However, the results of the study provide insight in the relative effectiveness of different degrees of variability in training, and can thus support the design of training programs.

2 Methods

2.1 Study population

From the 110 participants signing up for the study, 33 dropped out prematurely. One participant was removed after the first test because the experimenter noticed that this person had not been playing the game seriously (i.e. 31% drop out rate). Reasons for dropout were, among others: misjudging the amount of time needed for participation, bad internet connections, or disliking the game.

At the end, 76 (56.6% male) paid volunteers completed the training and the two tests for this study. On average the participants were 26.6 years old (SD=6.96; Range 18-45) and had a fair amount of gaming experience as rated on a 1-5 point scale ranging from 1 = "very little" to 5 = "very much" (M= 3.46, SD=0.90). Participants were assigned to one of the three groups by using a matching procedure based on sex, age and gaming experience. This resulted in 26 participants in the low variability group, 25 in the medium variability group, and 25 in the high variability group.

Participants received a compensation of 50 euros when they completed the experiment. To encourage their engagement, an additional prize of 50 euros was awarded to the participants who performed best of their groups in the tests. Participants were recruited by a brief storyline with images, medals, promotions, and awards they could win by participating in this study.

2.2 Sample size

A-priori power analysis was conducted to determine the necessary N for the study (*F*-test, repeated measures within-between interaction) using G*Power 3.1 (Faul, Erfelder, Buchner, & Lang, 2009). Parameters were set as follows: $\alpha = 0.05$, $1-\beta = 0.95$, medium effect size $f = 0.25$, comparing three groups and two measurements, correlations among repeated measure = 0.5, non-sphericity correction was set to one. The analysis showed that a minimum sample size of 66 was required to detect a medium-size effect.

2.3 Research tool

2.3.1 Online environment

The experiment ran on the online web-based tool 'The learning project' (in Dutch: "Het Leerproject", www.hetleerproject.nl). This tool has been developed within an earlier research program, and has been used for other training studies on personalized learning (Davidse, 2020)

(Van Mourik, 2020) (Van Dijk, 2020). Via the web-based tool, all of the materials for the experiment were presented to the participants (i.e., the introduction, all instructions, Space Fortress game sessions, feedback, and the questionnaires.) Personalized links to register and create an account on the website were sent out to the participants. Participants could individually plan the start of the training. Once started, programmed scripts ensured all tasks necessary to complete the study would load automatically.

2.3.2 The Space Fortress game

The research was performed with an adaptation of the serious game "Space Fortress" (SF), which has been specifically developed for research on the acquisition of complex skills (Frederikson & White, 1980). The game is similar to the arcade game Asteroids released by Atari. The visuals that were used were an improved version of the classic visuals, created in Unity by the Dutch research organizations NLR and TNO.

Figure 1 displays a screenshot of the game. The playing field is square, with a stationary (space) fortress in the center. Participants control a spaceship that flies within this square and they are instructed to fly around the Fortress in the middle (see) by accelerating forward (vehicle-frame), or rotating (yaw) left and right. Participants were advised to use their dominant hand on either the arrow keys (right), or the w, a, d keys (left). To decelerate, participants have to rotate the ship 180 degrees around, and accelerate in the opposite direction to the flight path. There was a maximum speed, but as there was no friction, there was no automatic deceleration. Except for the white letter 'd' identifying the mine type, the information presented at the bottom of the screen was irrelevant for the experiment, and participants were instructed to disregard this.

The fortress fires missiles at the ship every 5-6 seconds. Hits by these missiles should be avoided. Yet, for the purpose of the present study, hits received by the ship had no damaging effect but did give a visual cue and were logged as a flight indicator. A green hexagon serves as a reference of the most effective flight path to dodge hits. When exceeding the boundaries of the square playing field, the ship would appear on the opposite side moving with the same velocity and direction (i.e., "wraparound"). Furthermore, participants were instructed to destroy the Fortress using a certain procedure (i.e., after five shots/hits the Fortress becomes vulnerable and can be destroyed with a double shot).

Figure 1. Screenshot of the game with the ship (top), mine (left), fortress (centre), fortress counter (right of fortress), fortress missile (below near the ship) and mine type indicator (white d in bottom-centre).

Besides these basic tasks of flying, shooting, and missile avoiding, another task element was introduced: remembering and executing a procedure under time pressure. A mine would appear every 20 seconds on the edge of the playing field, moving towards the ship at a slower velocity than the ship could fly (see Figure 1 left side). While a mine was present, the fortress was invulnerable, yielding firing shots at the fortress to be pointless. The mine had to be dismantled using a specific procedure. There were five types of mines, each with their own procedure, which was indicated by the white letter in the center of the bottom yellow text (see Figure 1bottom, under “IFF”). The different mine types are listed in Table 1.

If an error were made in the procedure, there would be no feedback of success (see Table 1) as the indicated information did not change. Participants could then attempt to execute the procedure step correctly without having to start the procedure over from the beginning. If a mine was not successfully dismantled after 15 seconds, it self-destructed after a visual cue that was clearly different from the visual cue of dismantling the mine.

Table 1. Mine types and required procedures

Mine ID	Letter	Procedure step 1	Feedback	Procedure step 2	Feedback	Procedure step 3	Feedback
1	D	Double tap "5"	Letter turns green	Hold "8" for 3 seconds	Letter disappears	Tap "m"	Mine destroyed
2	d	Hold "j" for 3 seconds	Letter turns green	Tap: "7"	Letter disappears	Tap "m"	Mine destroyed
3	***d***	Tap: "vh"	Letter turns green	Hold "6" for 3 seconds	Letter disappears	Tap "m"	Mine destroyed
4*	D	Double tap "5"			Letter disappears	Tap "m"	Mine destroyed
5	None					Tap "m"	Mine destroyed

Note. *Mine type 4 was a variation of mine type 1, and used only in one of the tests to present a novel situation to the participants. For this mine, the letter immediately disappeared after step 1 of the procedure. This behavior indicates that the "m" was to be pressed immediately, as was also the case for the other (trained) mine types.

The participants were told that the game had the following, equally important objectives:

1. Destroying the fortress as often as possible in the game session.
2. Dodging fire by the fortress;
3. Dodging mines;
4. Dismantling the mines as quickly as possible.

2.3.3 Analogies of the experimental tasks with flying tasks

The skills required for SF have similarities with the skills pilots need in flying an aircraft. Research has shown that training in SF can improve certain skills of fighter pilots, such as control behavior and attention management under high workload (Gopher, Well, & Bareket, 1994). In the scope of the current experiment, we see the following parallels with flying:

- The basic task (flying and shooting) represents manual flying skills. These are psychomotor skills that can be performed with relatively little attention;

- The procedures to dismantle the mines are representative of executing flight procedures. In the repetitive test, responding to the mines represents the task of executing flight procedures in normal/expected situations. In the variable test, responding to the mines represents the task of executing memory items involved in dealing with unexpected situations, such as aircraft emergencies;
- The new mine type used in the tests, can be considered an unexpected event (emergency) that is not explicitly trained. This event requires adaptive skills, or diverging from known procedures.

2.4 Protocol

All participants took part in four parts: two training blocks (i.e., the basic skills training, and procedure training), and two tests. Each training blocks lasted about two hours, and the tests lasted about 30 minutes each. The direct test was performed immediately after the two training blocks were finished. The retention test took place one week later. Participants were instructed that both training blocks and the direct test needed to be completed within two days.

Each element of the game was explained and practiced incrementally in various game sessions. Thus, each training blocks consisted of several game sessions of 3-7 minutes in which the game was continuously played (the duration of the game sessions varied based on the task that was being trained, but was the same for every participant). Each session was preceded by a briefing text with instructions, and concluded by a debriefing text with feedback on the participant's performance.

2.4.1 Training of the basic task

After receiving information about the experiment, providing informed consent, and completing questionnaires, the participants started with the training of the basic flying and shooting skills, together designated the "basic task". This task was not part of the experimental manipulation, and was the same for all participants.

Table 2 describes the 16 basic training sessions.

Table 2. Basic training sessions

ID	Contents	Duration including instructions and feedback (minutes)
1	Free practice with controlling the ship	4
2	Exercise to stop the ship a number of times.	5
3	Exercise to catch appearing targets in order to fly around the fortress.	7
4	Repetition of 3.	7
5	Exercise to fly around the fortress without targets, but with instruction to not exceed the playing field borders.	7
6	Flying around the fortress while dodging fire by the fortress.	7
7	Shooting at the fortress with single shots. Fortress is not firing.	7
8	Repetition of 7.	7
9	Practicing the double shot on the fortress while the ship is stationary. Fortress is not firing.	3
10	Destroying the fortress while the fortress is not firing.	7
11	Repetition of 10.	7
12	Repetition of 10.	7
13	Destroying the fortress while dodging fire by the fortress.	7
14	Repetition of 13.	7
15	Destroying the fortress while the fortress is firing and avoiding mines that appear. No procedure is given yet.	7
16	Repetition of 15.	7

A break was suggested after session 7. Participants received feedback about their performance, such as the number of successful shots in session 7; targets caught in session 3 and 4; fortress destroyed in session 10-16; and fire dodged in session 6 and 13-16. Where possible, task performance was compared with preceding sessions. Improvements were rewarded with compliments, and decrements were accompanied by encouraging messages, with the objective to

keep participants motivated. All participants were rewarded for achievements with badges or promotions at fixed moments during the training. The total duration of playing the game was 103 minutes. With briefing, debriefing, and loading times, the maximum duration of the basic training was 120 minutes.

2.4.2 Training of (mine dismantling) procedures

After completing the basic task training, the participants started the training of the procedures for dismantling the mines, designated the "procedure task". For all groups the training of the procedure task consisted of 9 sessions of 5 minutes each, summing up to 45 minutes playing time. Including time for briefing, debriefing, and loading the game, this training block lasted for a maximum of 60 minutes. In each session, participants practiced dismantling one, two, or three different mine types (see Table 1) totaling 12 mines per game session. Before each session, the required procedures for the upcoming mine types were briefed. After each session, participants received feedback about their performance, i.e., the number of mines successfully destroyed, and the average speed of procedure execution, both compared to the previous session.

2.4.3 Design

The variation of practice in the procedures to be learned was manipulated between subjects. Three different patterns of variation in practicing the procedures were used, as shown in Table 3:

- Group 1 (Low variability, "Low-Var") practiced one dismantling procedure per training session, where the same procedure was repeated as much as possible. The practice was presented as much as possible in a "blocked" manner;
- Group 2 (Medium variability, "Med-Var") first practiced the same procedure in sessions like the Low Var group, but later practiced with variations, meaning that multiple mine types were appearing within sessions. The practice thus went from "blocked" to increasingly "random";
- Group 3 (High variability, "High-Var") practiced all three procedures in the same sessions immediately from the start. The practice was presented as much as possible in a "random" manner.

Table 3. Mine types presented in the procedure training

Session	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
	Mine types presented in each session								
Low-Var	1	1	1	2	2	2	3	3	3
Med-Var	1	2	1/2	3	1/3	2/3	1/2/3	1/2/3	1/2/3
High-Var	1/2/3	1/2/3	1/2/3	1/2/3	1/2/3	1/2/3	1/2/3	1/2/3	1/2/3

Only the pattern of variation was manipulated over training: not the overall exposure to mines. Thus, each group received the same number of mines, and the same types of mines. In training sessions that contained more than one mine type, the mine types were presented in random order. New mine types were introduced in the instructions before the session.

Since participants in the low-variability group received only one type of mine during a session, we anticipated an unwanted effect if they would discover that they do not need to process the letter presented at the bottom of the screen. This would have induced a different execution of the procedure as intended. To make sure that processing the letter remained a part of the procedure, an additional mine type was added to the training. This was mine type 5 (see Table 1), which appeared without a letter and had to be dismantled by only performing the last step common to each procedure. The ratio in which mine type 5 appeared compared to the other mine types was 1:4.

2.4.4 Tests

The Direct test followed immediately after the training, and the Retention test took place one week later. Both tests consisted of the same parts. Participants were reminded beforehand of the importance to perform well in the test; of the game objectives (quick procedure execution, destroying the fortress, and dodging shots and mines) and of the monetary reward. Each test session was preceded by a briefing about the upcoming mine types and necessary procedures. This was done to negate any potential advantage for participants who wrote the procedure down on paper to remember it better.

Both the Direct test and the Retention test consisted of the following three parts in the following order:

- **Repetitive (baseline) test**, in which only mine type 2 appeared. This first repetitive post-test was used as a baseline;
- **Variable test**, in which mine type 1, 2 and 3 appeared. Here, the adaptive skills of the participants were tested, as they had to switch between different procedures;
- **New test**, in which mine type 1, 2, 3 and 4 appeared. Here only the procedure performance of mine type 4 was analyzed as a measure of the participants' capability to transfer their learned procedure to deal with a new situation.

2.4.5 Hypotheses

Regarding the procedure task we expected to find the following ranking of the groups:

1. Repetitive test: Low-Var > Med-Var > High-Var
2. Variable/new mine test: Med-Var > Low-Var / High-Var

The explanation for 1) is that it was expected that training with little variation is best suited to develop routine skills, as tested in the Repetitive test. The explanation for 2) is that it was expected that variable skills, as tested in the Variable and New mine test, were acquired in an optimal manner in the Med-Var group, and less optimal in both the Low-Var and the High-Var groups.

Regarding the performance on the basic task, we expected to the find the following ranking of the groups across all tests: Low-Var > Med-Var > High-Var.

This is based on our assumption that the cognitive load of training the procedures would be lowest in the Low-Var group, leaving more resources available for further refining of their skills with regard to the basic task.

In sum, negative transfer of low variability training was expected on adaptive skills, and negative transfer of high variability training was expected on repetitive tasks, and on basic flying skills.

2.5 Dependent variables

2.5.1 Procedure task performance

The performance on the procedure task (mine dismantling) was measured by means of three objective variables:

- **Mine dismantling time:** the average time from the appearance of the mine until it was successfully dismantled. If a mine was not dismantled, the maximum time (15 seconds) was counted.
- **Procedure errors:** the average number of erroneous button presses per mine. This includes pressing the correct button in an incorrect manner (i.e. single tapping instead of double tapping, or holding it too briefly.
- **Procedure success rate:** the original intention was to use percentage of mines successfully destroyed in each test. However, many participants achieved a 100% success rate in tests. Therefore, procedure success could not be measured on an interval scale. A dichotomous measure was used instead, which reflects either "all mines in the session destroyed" or "at least one mine not destroyed".

2.5.2 Basic task performance

The flying and shooting performance was analyzed by means of three objective variables:

- **Fortress destroyed:** the average number of times the fortress was successfully destroyed in each session.
- **Fortress fire dodged:** the average number of shots by the fortress that were dodged.
- **Playing field limits exceeded:** the number of times the participant exceeded the playing field limits.

2.5.3 Subjective measures

To check whether groups were equal at the start of the experiment with respect to their motivation to participate, the Interest and Enjoyment subscale of the Intrinsic Motivation Inventory (Ryan, 1982) was administered, which consisted of seven items.

Mental demand was assessed multiple times during the study, after (1) the basic skills training, (2) the mine dismantling training, (3) the Direct test, and (4) the Retention test. A version of the Mental Demand subscale of the NASA Task Load Index (NASA-TLX); (Hart & Staveland, 1988) was used. A seven-item Likert scale was used to record the participant's answers.

2.6 Data analysis

The performance variables were compared between the groups in the repetitive, the variable, and new mine Direct and Retention tests (separately) using (between subject) Kruskal-Wallis for

non-parametric data. Binary measures are compared between the groups using (between-subject) Chi-squared tests.

The other dependent variables were compared between the groups for each measuring moment using one-way ANOVA analyses.

3 Results

3.1 Exclusion of participants

We carefully monitored the progress of the participants during the experiment, and we had to end the experiment for one participant after discovering that he/she was not observing our instructions (the space ship remained in a stationary position).

Data of six more participants were excluded from the analysis afterwards, because the amount of fire dodged by these participants was below five, whereas the average is around 37. This was also the case for one more participant in the Retention tests.

Dodging fire of the fortress is an indicator of good flying. A low score on this variable means that the participants were not flying around the fortress as instructed. Since the double-task of flying also affected procedure performance, a lack of effort for the flying task would give participants an unfair advantage on executing the procedures. Therefore, all performance variables (including procedure performance) were excluded from analysis for these participants.

After exclusion, the number of participants for the Direct test was 25 in the Low-Var group; 23 in the Med-Var group; and 22 in the High-Var group. For the Retention test, the numbers were 24 in Low-Var group, 23 in the Med-Var group, and 22 in the High-Var group.

3.2 Comparison of participant characteristics between the groups

To check that any differences in the performance were not a result of differences in the characteristics of the participants between the three groups, we performed two one-way Analyses of Variance (i.e., with age and gaming experience as dependent variables and condition as factor), and a Chi-square test (with gender as dependent variable and condition as factor). The descriptive statistics of the groups are presented in Table 4, showing small differences between the distributions of gender, age and gaming experience between the groups.

Table 4. Descriptives of the participant characteristics per group

	N	**Sex**		**Age**		**Gaming experience**	
		Male	Female	mean	SD	mean	SD
Low-Var	25	14	11	25.00	5.69	3.44	.92
Med-Var	23	14	9	27.70	8.23	3.61	1.03
High-Var	22	11	11	25.82	5.30	3.32	.84
Total	70	39	31	26.14	6.53	3.46	0.93

The results of the two one-way Analyses of Variance did not show significant differences between the conditions in Age, $F(2,67)= 1.06$, $p = 0.352$, and gaming experience, $F(2,67) = .551$, $p = 0.579$. The results of the Chi-square test did not show a significant difference in gender between the groups, $\chi2(2,67)= .54$, $p = 0.763$. Thus, the groups appear to be well balanced, as they are not significantly different in age, gaming experience, and distribution of males and females.

3.3 Effects of variability of practice on the development of the procedural skill

3.3.1 Mine dismantling time

Figure 2 shows the average time needed to execute the correct procedure for the three groups in all sessions during the training (white area), the Direct test (green area), and the Retention test (blue area). Note that performance in the new mine test concerns only the presentation of the new mine (mine type 4). The next sections will discuss the results in this figure for the training and test sessions separately.

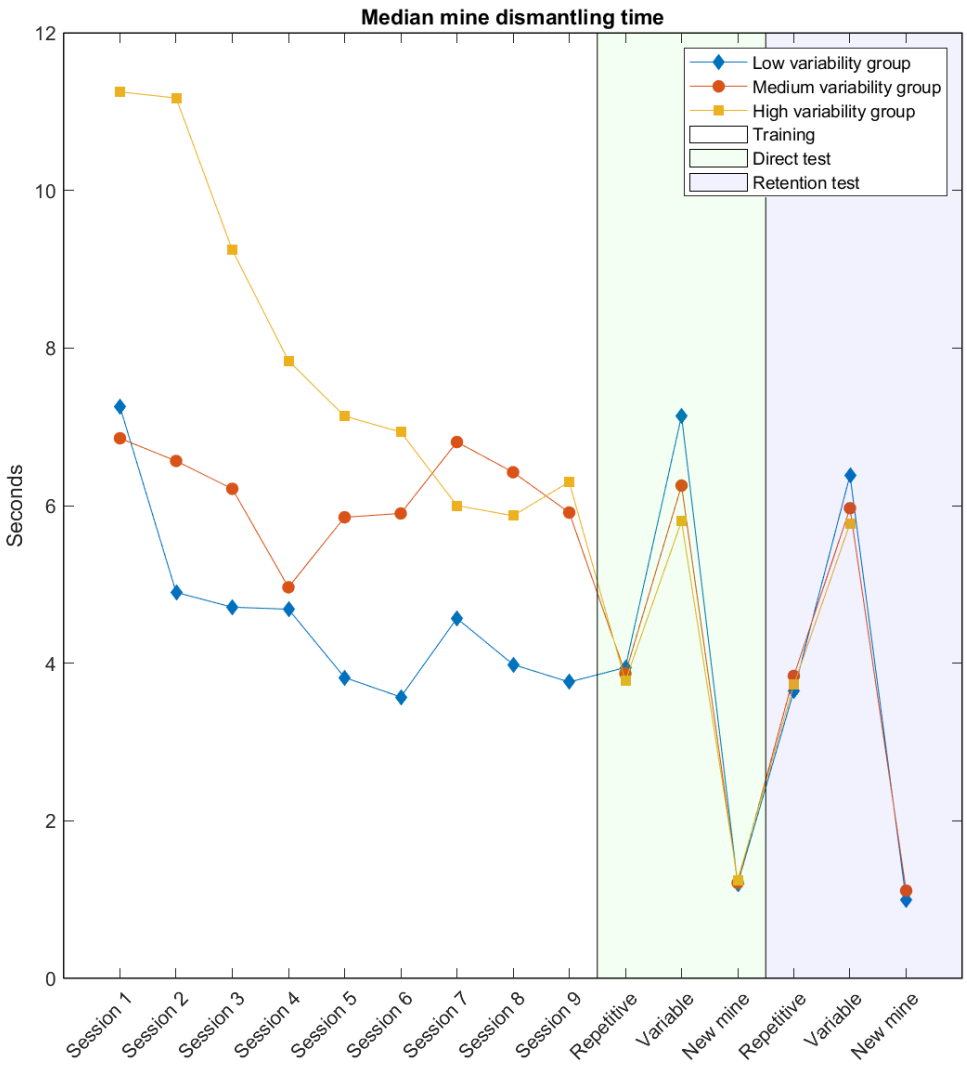

Figure 2. The median time to correct procedure execution for the three groups during the training sessions and the test sessions.

3.3.1.1 Training

As can be seen in the white area of Figure 2, the High-Var group initially took very long to execute the procedures, but as training progressed, the median time gradually decreased, reflecting a classical learning curve. In the last three training sessions (which featured three mine types for the Med-Var and High-Var groups), the performance of the Med-Var and High-Var groups was very similar.

Figure 2 also shows that the participants in the Low-Var group generally reached shorter dismantle times than both other groups in the training. After nine training sessions, their median time needed to dismantle a mine amounted to 4.00 sec. The figure also shows that determining and executing the correct procedure for each of the three types of mines (see last training session of the Med Var and High Var groups) took participants about two seconds longer than simply responding to one type of mine (last training session of the Low-Var group).

3.3.1.2 Repetitive test

The analysis of the Repetitive test did not show significant differences between groups ($p = 0.980$). This shows that all groups performed equally on a single procedure task, irrespective of whether they were trained on one procedure per session (low-var group), or on multiple procedures per session (med-var en high-var groups).

3.3.1.3 Variable test

The analysis of the Variable test in the Direct test showed a marginally significant effect of Training Variability for the average dismantle time, $H(2) = 5.20$, $p = 0.074$. As can be seen in Figure 3, the average time was longest in the Low-Var group. This group was significantly slower than the High-Var group, $U = 166$, $p = 0.020$, but not compared to the Med-Var group, $U = 223$, $p = 0.183$. There was no significant difference in the retention test.

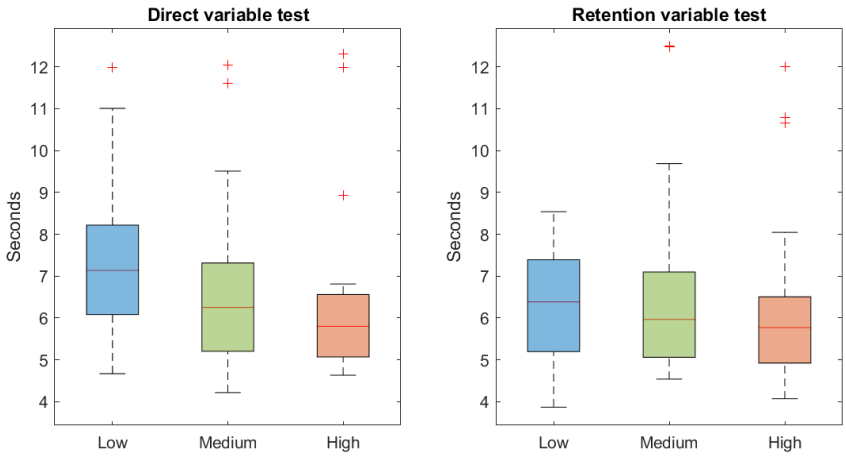

Figure 3. Tukey Boxplots of the mine dismantling time in the direct variable test and the retention variable test.

3.3.1.4 New mine test

When responding to a new type of mine in the Direct test, there was no significant difference between the groups in the average time needed to execute the correct procedure, $p = 0.814$, also not when only considering the first encounter of the new mine, $p = 0.678$. This was the same for the new mine test in the Retention test.

3.3.2 Procedure errors

3.3.2.1 Repetitive test

The Low-Var group seemed to make the least procedure errors, but this effect was not significant, $p = 0.209$. A similar but non-significant pattern was seen in the Retention tests see Figure 4.

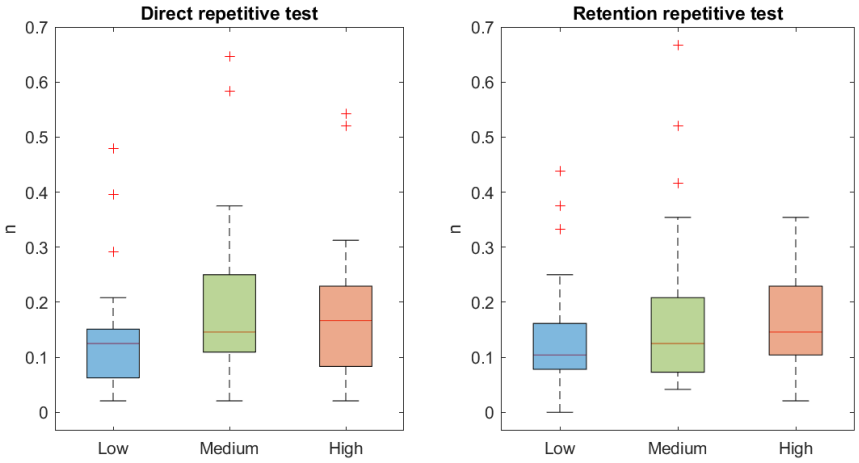

Figure 4. Tukey boxplots of the number of procedure errors (per session) in the direct and retention repetitive test.

3.3.2.2 Variable test

Looking at the number of procedure errors in the Variable test as part of the Direct test, there was no significant difference between the groups, p = 0.701. However, as shown in Figure 5, there was a significant difference in average number of errors in the first procedure step, p = 0.024, indicating that the Low-Var group made more errors than the Med-Var, U = 171, p = 0.016, and the High-Var group, U = 174, p = 0.031. There was no significant difference in the retention test.

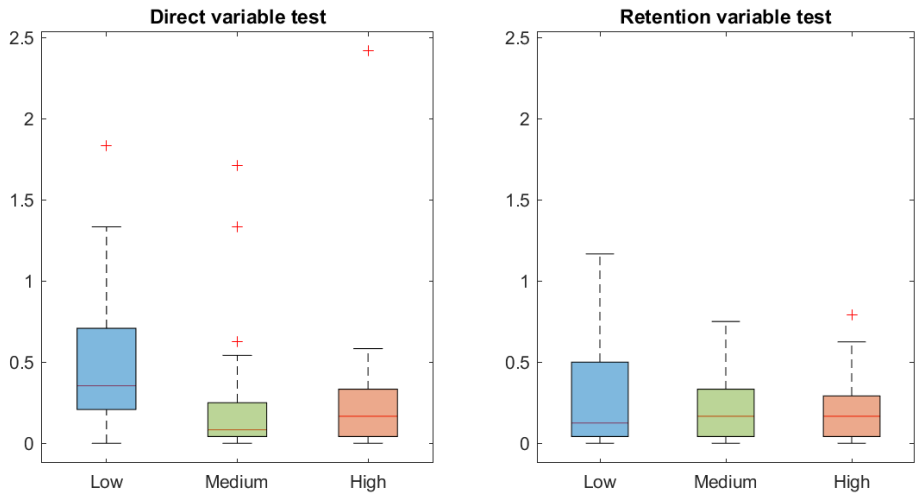

Figure 5. Tukey boxplots of the average number of procedure errors per mine in the first procedure step in the variable test as part of the direct test and the retention test.

3.3.2.3 New mine test

As shown in Figure 6 there was a significant difference between the groups concerning the average number of procedure errors in the Direct New mine test, H(2) = 6.02, p = 0.049. The High-Var group made more procedure errors than the Low-Var group, p = 0.017. There were no significant differences in the Retention test, p = 0.526.

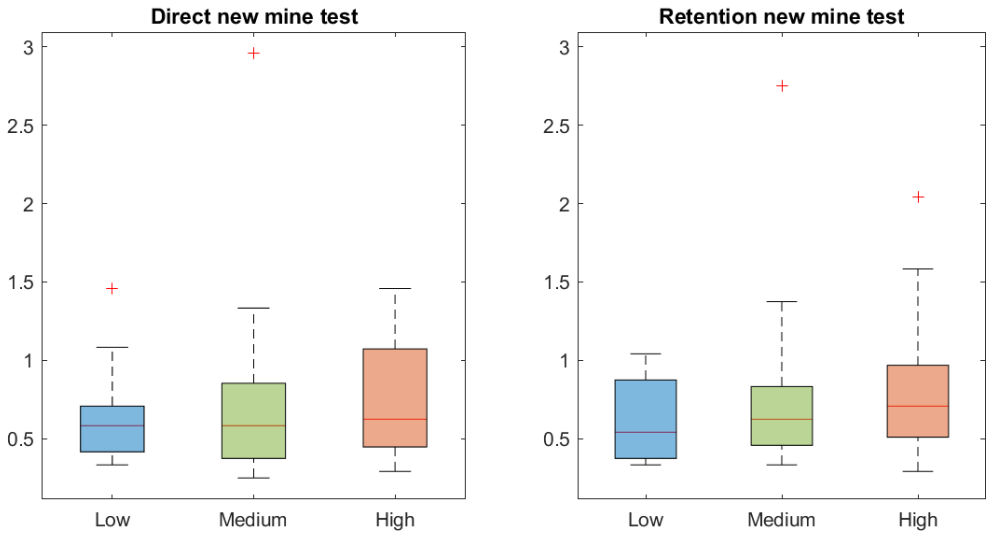

Figure 6. Tukey boxplots of the average number of procedure errors per mine (only new mine type) in the new mine test as part of the direct test and the retention test.

3.3.2.4 Procedure success rate

For each test, Figure 7 shows the proportion of participants in each group who successfully dismantled all mines. Statistical analysis showed no effect of Training Variability, as will be explained in the next sections for each test separately.

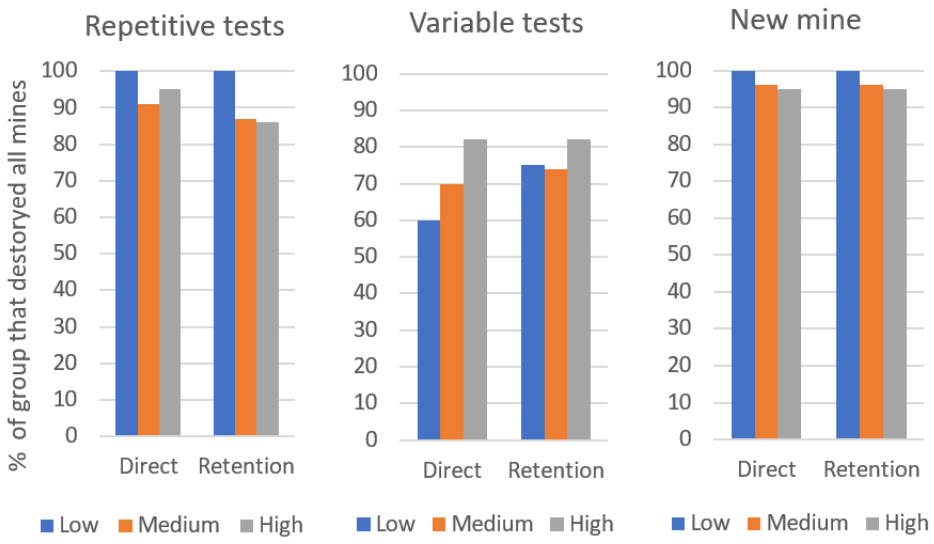

Figure 7. The proportion (percentage) of participants in each group who successfully destroyed all mines in the tests.

3.3.2.5 Repetitive test

In the Direct Repetitive test (left-hand side of left plot in Figure 7), two participants in the Med-Var group, and one participant in the High-Var group failed to destroy all mines, whereas none of the participants in the Low-Var group did. This difference was not significant, $p = 0.331$. A similar trend was observed in the Repetitive test as part of the Retention test one week later (right-hand side of left plot in Figure 7), where three participants in the Med-Var group, and three in the High-Var group failed to dismantle all mines. This difference was also not significant, $p = 0.164$.

3.3.2.6 Variable test

Looking at the Direct Variable test (left-hand side of middle plot in Figure 7), the Low-Var group was less likely to destroy all mines, whereas the High-Var group performed best, but this difference was not significant, $p = 0.265$. In the Repetitive test as part of the Retention test (right-hand side of middle plot in Figure 7, there were similar patterns, but also without significant effects.

3.3.2.7 New mine test

The new mine was not always destroyed in the Direct test by one participant in the Med-Var group, and by one participant in the High-Var group. In the Retention test, this was only the case for one participant in the Med-Var group. These effects were not significant.

3.4 Effects of variability of practice on the development of the basic task skills

The measures for basic task performance did not significantly differ between the different conditions of training variability (see Table 5). For the Direct test, the p-values were: Fortress destroyed, $p = 0.502$; Fire dodged, $p = 0.911$; Exceeded playing field, $p = 0.664$. For the Retention test, these values were: Fortress destroyed, $p = 0.540$; Fire dodged, $p = 0.968$; Exceeded playing field, $p = 0.609$.

Table 5. The median performance scores (per session) of the basic task performance task

		Low Var	**Med Var**	**High Var**	**Low Var**	**Med Var**	**High Var**
		Direct tests (median)			**Retention tests (median)**		
Destroyed Fort (n/s)		.134	.108	.125	.114	.103	.117
Dodged fire (n)		39.5	37.3	37.6	39.6	38.5	39.5
Exceeded playing field (n)		5.50	7.17	8.42	6.17	6.17	6.33
Note: The basic task performance was analyzed over all Direct tests and Retention tests together.							

3.5 Effects of variability of practice on subjective mental demand during the training and the tests

3.5.1 Training

The mental demand scores for the training are shown in Figure 8. There was no significant effect for the subjective mental demand for the training of the basic skills, $F(2,67) = 1.23$, $p = 0.298$. This further confirms that the groups were likely well balanced in terms of gaming skills. For the training of the procedures (mine dismantling) there was a significant effect for subjective mental demand, $F(2,67) = 4.75$, $p = 0.012$. Post-hoc comparisons showed that the Med-Var group reported higher mental demand than the Low-Var group, $p = 0.019$, and higher mental demand than the High-Var group, $p = 0.005$.

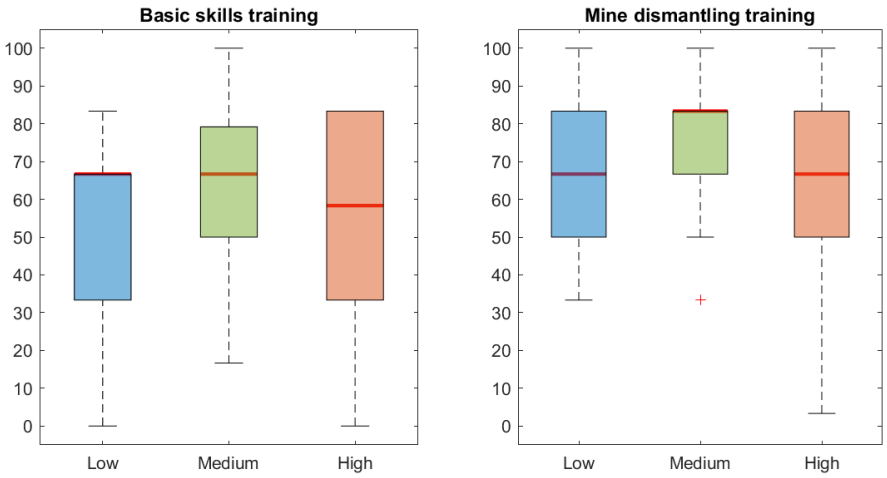

Figure 8. Subjective mental demand (Nasa TLX) in the basic skills training and in the mine dismantling training.

3.5.2 Direct and Retention tests

The mental demand scores during the tests are shown in Figure 9. There was a significant effect in the Direct test, $F(2,67) = 3.80$, $p = 0.027$, but not in the Retention test, $p = 0.541$. In the Direct test, the High-Var group reported lower mental demand than the Low-Var group, $p = 0.027$, and lower mental demand than the Med-Var group, $p = .014$.

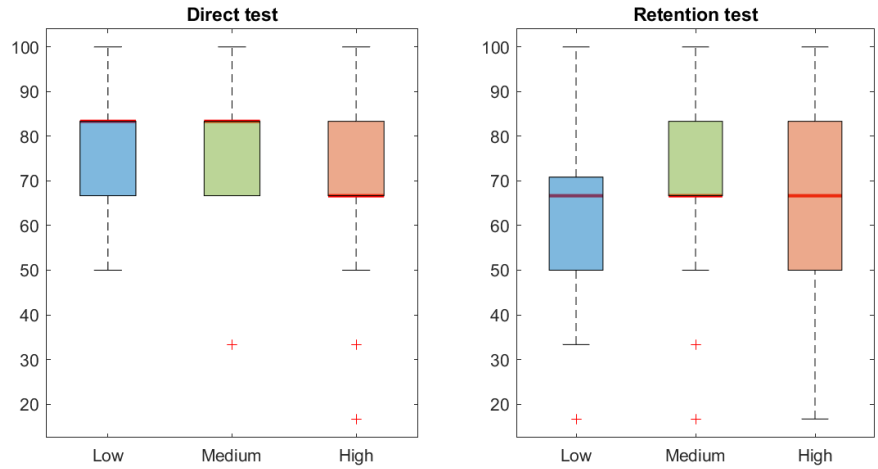

Figure 9. Subjective mental demand (Nasa TLX) in the direct test and retention test.

3.6 Effects of variability of practice on interest and enjoyment during the training

For the IMI Interest/Enjoyment subscale scores over the training, there was no significant effect, $F(2,67) = .723$, $p = .489$. Thus, there was no evidence that some groups enjoyed their training more than other groups. The mean score on the scale was 4.09, which is almost the midpoint (i.e. 4).

4 Discussion

Overall, we found evidence of negative transfer of training for low- as well as high-variability of practice. Both training paradigms hampered performance on the procedure task, but for different situations in which the task was tested. The significant effects only showed in the Direct test, not in the Retention test. The next sections will therefore discuss the main findings of only the Direct test in relation to our hypotheses listed in section 2.4.5.

4.1 Procedure task

4.1.1 Repetitive test

The Procedure task in the Repetitive test requires *routine* expertise. We assumed this would be developed best by low variability training and expected superior performance in the Low-Var group, and inferior performance in the High-Var group. However, the results of this test did not show differences between the groups for any of the tested measures (dismantle time, procedure errors, or success rate). Hence, the expected negative effects of high variability of practice on routine skills were not confirmed, nor were the expected positive effects of low variability of practice. It is not clear why the results did not corroborate our expectations, and the findings from literature (Bannert, 2002). Overviewing the results obtained, we feel that the tasks to be learned might have been too easy for participants, thus reducing the likelihood of demonstrating benefits of repetitive training.

4.1.2 Variable test

Since the Procedure task in the Variable test requires adaptive expertise, we expected superior performance by the Med-Var group, and inferior performance by the Low-Var group, when compared to the Med-Var group. Indeed, results confirmed that the Low-Var group took longer to dismantle the mines, and made more errors in the first procedure step, compared to the other groups. What might be the reason for Low-Var participants to make errors on the first step? To understand this finding, it is useful to make a distinction between selecting a procedure, and executing one. Participants in the Low-Var group had practiced the execution of the procedure extensively. However, before the test, they never had to select the procedure that fitted the circumstances. Because of this, they were more likely to select an inappropriate procedure than the participants in the other groups were. The High-Var group, in contrast, performed equally well as the Med-Var group. So, no evidence was found that acquiring the ability to respond adaptively to frequently changing task situations is hampered by high variability during training.

The results support the hypothesis that repetitive practice during the training of procedures has a negative effect on the acquisition of adaptive skills.

Thus, repetitive tests to check whether learners have acquired the ability to execute a particular procedure do not reveal a learner's ability to adapt to situations that are different from the trained situations. This conclusion has consequences for certification of professionals. For example, after following a refresher training, pilots may perform adequately when a test is being used that presents predictable situations only. However, this does not automatically reveal their status with respect to adaptive skills. The results demonstrate that the application of trained procedures should not be tested only in situations that are equal to the conditions during training, but also in simulations of situations that are new and unpredictable.

4.1.3 New procedure test

Similar to the Variable test, the New mine test also involves adaptive expertise. The High-Var group made most errors in this test, and performed significantly worse than the Low-Var group. This is consistent with our hypothesis that training under high variability conditions has a negative impact on acquiring adaptive expertise. How can this be interpreted? Again, it may have to do with the distinction between selecting and executing a procedure. As the High-Var group never repetitively practiced a procedure, they were likely to be more focused on selecting the correct procedure than on memorizing the connection between the presented information (i.e., the letter turning green and then disappearing) and the sequence of actions. Thus, when they were presented with a new situation in the new procedure test, they experienced more difficulty at deducing the appropriate response than participants of the other groups.

One hypothesis was that the Low-Var group would perform poorly on this *new procedure test*, as their training was designed to acquire a solid expertise in procedure execution, not for developing adaptive expertise. It is therefore surprising that this group actually performed best. It may be that this unexpected outcome is related to the chosen procedure to dismantle the new mine (type 4). The procedure for this mine was designed to be similar to mine type 1. The first step of the procedure was the same (the same letter), but in the new mine procedure, the second step was skipped. However, during the training sessions there was also a "decoy" mine (type 5), which was presented without a corresponding letter. Dismantling this mine only required the last step of each procedure, which was pressing "m" (see Table 1). Mine type 5 was randomly presented during the training sessions to make sure that all groups looked at the presented letter when the mine appeared. Thus, unintentionally, the design of the experiment caused the Low-Var group to train the two required procedure steps of the mine in the *new procedure test* in a

variable manner, as they had responded during three training sessions to both mine type 1 and mine type 5.

But even when taking into account this unintentional side-effect of the experimental design, it is interesting to note that the Low-Var group performed best to the new mine, and significantly better than the High-Var group. After all, each group had practiced the necessary parts of the procedure for the new mine an equal amount of times. Perhaps the Low-Var group's repetitive training sessions, which only focused on varying mine type 1 and 5, gave them an advantage over the other groups. It would mean that the training of procedures is most effective when variability of practice focuses on the elements that differentiate situations, and use this to learn the matching sequences of responses.

4.1.4 Retention

The overall performance level achieved at completing the training and tests seemed to be retained in the retention tests (see Figure 2). However, the differences between the groups dissipated. Perhaps the Direct test negated the group differences in the Retention test, as the Direct test itself may have provided each group with some repetitive as well as variable training.

4.2 Basic task

In contrast to our hypothesis that the repetitive training of the Low-Var group would result in a better basic skill level compared to the Med-Var and the High-Var groups, we found no significant differences between the groups for any of the measures. In other words, the results of this experiment indicate that variability in practice does not hinder pilots in developing and refining basic skills.

4.3 Limitations and lessons learned

4.3.1 Difficulty of the task during the training and tests

Several outcomes of the study indicate that the level of variability during training may have been insufficient to demonstrate substantial effects on the acquisition of basic skills. First, the High-Var group reported a similar mental load when training to dismantle the mines as the Low-Var group. Second, inspection of Figure 2 suggests that the performance in executing the procedures of the Med-Var and High-Var group converged in the last three training sessions, which should not be the case if the variation for the High-Var group had been "excessive." This convergence of performance was already suspected after pilot testing, which had led us to reduce the training time by three sessions for all groups. Still, the many repetitions of variable sessions may have

induced too much repetitiveness for the High-Var group. Perhaps, including more, or more complex procedures, would have amplified any negative effects of high variability on basic task performance (and subjective mental demand).

Thus, the study design could be improved by presenting more mine types to the High-Var group, which are not included in the test. This would increase the training time, although training time on the mines that are tested can be kept the same as in other training groups. A second improvement could be to make the required procedures more complex, with more steps or even decision trees. This would make the materials more similar to procedures used in actual professions, although it is likely that more training time would be required in the study. Nevertheless, such improvements may be required to increase the contrast between training conditions. The required procedure to respond to a new mine could then also be made more difficult, since almost all participants were able to figure out what the appropriate procedure was (see Figure 7).

4.3.2 Performing the experiment at home

To circumvent issues related to Covid-19, the experiment was administered over the internet, so that participants could do the training and tests from home. This probably decreased the threshold to participate, as a sufficient number of participants was found quickly and easily. However, it also gave experimenters limited insight and control over participant effort and concentration. During data analysis, some cases were excluded from the analysis when we suspected low effort on parts of the task. However, it is not possible to identify all these cases with 100% accuracy and certainty. Although we attempted to boost participant effort and motivation by introducing some achievements during the training, a brief storyline with pictures, and a monetary reward for test performance, participants still scored around the scale midpoint on Interest/Enjoyment for the training, which is in our experience somewhat low for an experimental training. Performing the tasks at the laboratory under supervision of an experimenter would have been better as personal attention may instill a better sense of importance as well as understanding of the tasks.

Following the training and tests from home allowed the participants to use certain loopholes, like, for example, writing the procedures down to use as a reference. Although executing the procedures using a paper reference would likely impair their performance compared to executing them more quickly from memory, this may still have distorted the performance measures in the test. In addition, we controlled for potential benefits from writing down the procedures by presenting the necessary procedures to each participant before each test. However, this likely made the tests less challenging than they could be.

4.4 Recommendations

The results of the experiment show that repetitive training of procedures is not the best approach to achieve adaptive performance. If professionals need to be able to adapt flexibly to new and unexpected situations, it is better to adopt a training schema with a certain level of variability.

The results also suggest that including a high level of variability in training hinders the flexible application of procedures in new situations (i.e., leading to errors as diverging from procedures, or erroneously combining parts of different procedures). An implication of these findings for the training of pilots is that repetitive training of emergency flight procedures should also be trained in a repetitive manner. This supports a solid mastery of skill when the circumstances requires its execution. It may be wise to use repetitive training in earlier stages of skill acquisition, to be followed by practice in more variable and unpredictable scenarios. This supports the development of a pilot's competency to detect variations and possible anomalies in a situation, which may require the pilot to adapt their response.

How much repetition or variability is appropriate for a specific flight training depends on the relative requirement, or importance of routine skills versus adaptive skills. Many flight tasks consist of routine skills. However, this also means that after completing a training, these routine skills are being further developed with each execution during operational practice. Recurrent training is focused specifically on routine skills. Adaptive skills are required in unpredictable or new situations, which are rather infrequent. However, the appropriate response is likely to be of critical importance. To determine which (aspects of) situations should be varied to develop sufficient adaptive expertise, it is important to analyze which situations may occur that are difficult to distinguish from each other; which of the situations require highly specific responses; and which of the situations require information that is known to be often overlooked. By focusing the variability in training on these determined specific aspects, training time can be used efficiently to develop the adaptive expertise that is specifically needed for the task and the job. In other words, training then focusses on strengthening the most important connections between information, information processing, and response.

5 Acknowledgements

This work was performed under contract ‘31161452 TNO-Phase II Negative training RQ1’ of the Aviation Directorate of the Dutch Ministry of Infrastructure and Water Management (MinIenW).

6 References

Alexander, A. L., Brunyé, T., Sidman, J., & Weil, S. A. (2005). *From gaming to training: a review of studies on fidelity, immersion, presence, and buy-in and their effects on transfer in PC-based simulations and games.* DARWARS Training Impact Group. doi:10.1016/j.athoracsur.2004.02.012

Arthur, W., Bennett Jr, W., Stanush, P., & McNelly, T. (1998). Factors that influence skill decay and retention: a quantitative review and analysis. *Human Performance 10.1207/s15327043hup1101_3, 11*(1), 57–101. doi:10.1207/s15327043hup1101_3

Baldwin, T. T., & Ford, J. K. (n.d.). Transfer of training: a review and directions for future research. *Personnel Psychology, 41*, 63-105. doi:10.1111/j.1744-6570.1988.tb00632.x

Bannert, M. (n.d.). Managing cognitive load - recent trends in cognitive load theory. *Learning and instruction, 12*(1), 139-146.

Borgvall, J., & Nählinder, S. (2008). *Transfer of training in military aviation.* IMTR-International Mission Training View project. doi:10.13140/RG.2.1.5087.7280

Burke, L. A. (1997). Improving positive transfer: a test of relapse prevention training on transfer outcomes. *Human Resource Development Quarterly, 8*, 115-128. doi:10.1002/hrdg.3920080204

Davidse, J. (2020). *The effects of guidance in self-regulated learning on learning progress and performance.* Master Thesis, Universiteit Utrecht.

Faul, F., Erfelder, E., Buchner, A., & Lang, A. (2009). Statistical power analyses using G*Power 3.1: tests for correlation and regression analyses. *Behavior Research Methods, 41*, 1149-1160.

Ford, J. K., & Schmidt, A. A. (2000). Emergency response training: strategies for enhancing real-world performance. *Journal of Hazardous Materials, 25*(2-3), 192-215.

Gopher, D., Well, M., & Bareket, T. (1994). Transfer of skill from a computer game trainer to flight. *Human Factors, 36*(3), 387-405. doi:10.1177/001872089403600301

Hart, S. G., & Staveland, L. E. (1988). Development of NASA-TLX (task load index): results of empirical and theoretical research. *Advances in Psychology, 52*, 139-183.

Landman, A., van Oorschot, P., van Paassen, M. M., Groen , E. L., Bronkhorst, A. W., & Mulder, M. (2018). Training pilots for unexpected events: a simulator study on the advantage of unpredictable and variable scenarios. *Human Factors, 60*(6), 793-805.

Pennings, H. M., Oprins, E., Schoevers, E., & Groen , E. L. (2019). *Current insights in negative transfer of training.* TNO report 2019 R11747.

Ryan, R. M. (1982). Control and information in the intrapersonal sphere: an extension of cognitive evaluation theory. *Journal of Personality and Social Psychology, 43*, 450-461.

Van Dijk, C. (2020). *Standaard Leertraject voor Space Fortress.* Master Thesis, Universiteit Nijmegen.

Van Merriënboer, J. G., & Kirschner, P. (2007). *Ten steps to complex learning: a systematic approach to four-component instructional design.* London: Lawrence Erlbaum Associates, Publishers. ISBN: 9780805857931.

Van Merriënboer, J. J., Kester, L., & Paas, F. (2006). Teaching complex rather than simple tasks: balancing the intrinsic and germance load ot enhance transfer of learning. *Applied Cognitive Psychology: The Official Journal of the Society for Applied Research in Memory and Cognition, 20*(3), 343-352.

Van Mourik, D. (2020). *The effects of scaffolding and feedback adaptive to the characteristics of the learner on learning progress and performance for the purpose of personalized learning.* Master Thesis, Open Universiteit Heerlen.

Vermulst, A., & Gerris, J. (2005). *QBF: Quick big five persoonlijkheidstest handleiding [quick big five personality test manual]*. Leeuwarden, The Netherlands: LDC Publications.

Woltz, D. J., Gardner, M. K., & Bell, B. G. (2000). Negative transfer errors in sequential cognitive skills: strong-but-wrong sequence application. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 26*(3), 601-25.