Training Pilots for Unexpected Events: A Simulator Study on the Advantage of Unpredictable and Variable Scenarios

Annemarie Landman, TNO Soesterberg, The Netherlands, **Peter van Oorschot**, **M. M. (René) van Paassen**, Delft University of Technology, The Netherlands, **Eric L. Groen**, **Adelbert W. Bronkhorst**, TNO Soesterberg, The Netherlands, and **Max Mulder**, Delft University of Technology, The Netherlands

Objective: This study tested whether simulator-based training of pilot responses to unexpected or novel events can be improved by including unpredictability and variability in training scenarios.

Background: Current regulations allow for highly predictable and invariable training, which may not be sufficient to prepare pilots for unexpected or novel situations in-flight. Training for surprise will become mandatory in the near future.

Method: Using an aircraft model largely unfamiliar to the participants, one group of 10 pilots (the unpredictable and variable [U/V] group) practiced responses to controllability issues in a relatively U/V manner. A control group of another 10 pilots practiced the same failures in a highly predictable and invariable manner. After the practice, performance of all pilots was tested in a surprise scenario, in which the pilots had to apply the learned knowledge. To control for surprise habituation and familiarization with the controls, two control tests were included.

Results: Whereas the U/V group required more time than the control group to identify failures during the practice, the results indicated superior understanding and performance in the U/V group as compared to the control group in the surprise test. There were no significant differences between the groups in surprise or performance in the control tests.

Conclusion: Given the results, we conclude that organizing pilot training in a more U/V way improves transfer of training to unexpected situations in-flight.

Application: The outcomes suggest that the inclusion of U/V simulator training scenarios is important when training pilots for unexpected situations.

Keywords: flight simulation, mental models, surprise, startle, training

Address correspondence to Annemarie Landman, Section of Control and Operations, Faculty of Aerospace Engineering, Delft University of Technology, Kluyverweg 1, 2629 HS, Delft, Netherlands; e-mail: h.m.landman@tudelft.nl.

HUMAN FACTORS
Vol. 60, No. 6, September 2018, pp. 793–805
DOI: 10.1177/0018720818779928

INTRODUCTION

According to regulations, the initial and recurrent type-rating checks for pilots must involve standardized and therefore relatively predictable scenarios. The training sessions for these checks can be organized in similarly predictable format (Bureau d'Enquêtes et d'Analyses, 2012; Casner, Geven, & Williams, 2013). However, research suggests that skills taught in this manner are "brittle" instead of adaptive (Casner et al., 2013; Landman, Groen, van Paassen, Bronkhorst, & Mulder, 2017b). That is, they transfer well to predictable situations like the tests, but they may not hold up in emergency situations, which are typically novel and unexpected. For this reason, many aviation companies look for ways to extend on the minimally required training. Aviation safety organizations have recognized the benefits of such extensions and offer ways to formalize them (see, e.g., Federal Aviation Administration, 1990).

Unfortunately, it is impossible to train pilots for every conceivable abnormal situation. However, pilots can be trained for common factors in such situations. One of these factors is the presence of startle and surprise, the training for which will become mandatory in the near future (European Aviation Safety Agency, 2015; Federal Aviation Administration, 2015; International Civil Aviation Organization, 2013). Coping with surprise requires effective sensemaking activities, which involves the search for a structured representation, or "frame" of the situation. This frame is used to direct attention, interpret information, and make sense of what is going on (see Klein, Phillips, Rall, & Peluso, 2007; Neisser, 1976; Rankin, Woltjer, & Field, 2016; Zhang, Soergel, Klavans, & Oard, 2008). Sensemaking includes, for instance,

TABLE 1: Characteristics of the Participants

	Group, *M* (*SD*)			
	U/V	Control	Δ	*p*
Age, years	41.3 (9.0)	41.5 (9.3)	0.2	.961
Employed as pilot, years	17.2 (8.8)	16.4 (7.3)	0.8	.827
Flight hours	9,311 (6,352)	7,571 (4,590)	1,740	.491

Note. U/V = unpredictable and variable.

seeking information on the instruments, reasoning, or testing out hypotheses. Such activities are particularly difficult when a pilot is startled (Eysenck, Derakshan, Santos, & Calvo, 2007; Landman et al., 2017a) or fatigued (Caldwell, 1997). If an appropriate frame cannot be found the perceived events may lose their meaning and coherence, which may increase stress even further and prevent effective troubleshooting.

Since predictable scenarios require very little sensemaking, they can be expected to be insufficient training to deal with surprise. A simple means to increase unpredictability would be to deny (certain) information about upcoming events. Second, aspects of a problem can vary among practice iterations, making each scenario somewhat different (Paas & Van Merriënboer, 1994; Van Merriënboer, Clark, & De Croock, 2002). Variable practice is thought to enhance a trainee's recognition of relevant features and rules, since the irrelevant features are different in each practice iteration (Shapiro & Schmidt, 1982). The resultant skills and knowledge are therefore better applicable in situations that are not explicitly practiced (i.e., generalization; Carbonell, Stalmeijer, Könings, Segers, & van Merriënboer, 2014). Finally, different types of problems can be intermixed (and spaced out) so that trainees cannot assume that the same solution is applicable in each next practice iteration. This is thought to improve their ability to pair the appropriate solution to the problem (Lee & Magill, 1985). Such methods are known as "mixed review" in math education (Rohrer, 2009) or "contextual interference" in perceptual-motor learning literature (e.g., Brady, 2004).

In the current simulator study, we investigated whether arranging simulator training in a more unpredictable and variable (U/V) manner improves pilot performance in a surprise test. The test required pilots to apply the learned knowledge and skills in a demanding and partly novel situation. Based on the aforementioned literature, it was hypothesized that U/V practice would improve the pilots' performance in this test, due to a better understanding of the events and the principles involved.

METHOD

Participants

Twenty participating airline pilots were randomly assigned to either a U/V practice group or a control group, unless when the groups became unbalanced with regard to the variables listed in Tables 1 and 2. Thus, two type-rating instructors and one pilot with extended light multiengine piston flying experience (i.e., CS-23/FAR part 23) were reassigned to the control group. All pilots, except one in the U/V group, indicated that they had at least 25 hours of light multiengine piston flying experience. Pilots were mainly recruited from a single airline company. Eight pilots from this company were in the U/V group and seven in the control group. This study complied with the tenets of the Declaration of Helsinki, and informed consent was obtained from each participant.

Apparatus

The experiment was performed in the Simona Research Simulator of the Delft University of Technology in the Netherlands, which features a 6–degrees of freedom hydraulic hexapod motion system and a collimated 180° (horizontal) × 40° (vertical) field-of-view display system. For outside visuals, the FlightGear open-source flight simulator was used. Standard washout filters were

TABLE 2: Characteristics of the Participants (Continued)

	Group, *n*	
	U/V	Control
Extended MEP experience, >50 hours	2	1
Type rating instructors	2	2
Rank		
Captains	4	5
First officers	6	5
Sex		
Men	9	9
Women	1	1

Note. MEP = multiengine piston; U/V = unpredictable and variable.

used for motion cueing (Reid & Nahon, 1988). A piston aircraft engine sound was played in mono over the pilots' headphones. Audio pitch was coupled to engine rpm and volume to torque. The cockpit mock-up was styled after a jet airliner and featured a B747-style primary flight display and a Cessna Citation–style engine display. Controls consisted of a right-hand sidestick with pitch trim control, rudder pedals with force feedback, and thrust, flap, and gear levers. A nonlinear aerodynamic model was used from a light twin-propeller aircraft: the Piper PA-34 Seneca III (De Muynck & Hesse, 1990; Koolstra, Herman, & Mulder, 2015). The aircraft model has certain properties and failure options, which allows for the development of challenging flying tasks. The airflow over the wing of each propeller induces extra lift, which causes a roll moment as well as a yaw moment in case of asymmetric thrust. At low speed, the moments generated by asymmetric thrust will exceed the maximum obtainable opposite moments generated by the control surfaces.

Tasks

The design of the experiment is illustrated in Figure 1. Each session listed in the figure is described in turn.

Briefing and familiarization. Pilots were informed that they would perform a number of scenarios and respond to malfunctions. They were instructed to complete the task as given (e.g., perform a landing, fly a circuit), unless a crash was imminent. Air traffic control communication and checklists were not included. Pilots were instructed to call out any problem as soon as they noticed it. They were informed about the required settings: a circuit would need to be flown at 1,000 ft with a speed of 130 kt and a power setting of approximately 42 N·m torque. Full throttle was used during takeoff. A flaps setting of 25° was required only during landing. The speed at rotate was 80 kt; the optimal rate of climb (V_2) was 92 kt; the approach speed was 85 kt; and the minimum control speed with a single engine was around 80 kt. Pilots were reminded of the settings if they diverged from them during the familiarization and practice. Wind direction and strength (light or moderate) information was provided verbally before each run and indicated by a wind sock located next to the runway. Following the briefing, pilots were seated in the simulator and instructed to fly two familiarization circuits: one without wind and one with light crosswind.

Practice session. Pilots then practiced with managing asymmetric thrust by performing six takeoff runs with a single-engine failure, four flyby runs with a rudder failure, and four flyby runs with a single-engine failure (described later). Aspects of the runs were varied in the U/V group only (see Table 3). The runs were presented in blocks of two (e.g., two takeoff runs; see Table 4). These blocks were alternated in the U/V group only. Before the first run of each block, pilots in the U/V group were merely informed that a malfunction would occur. Between the first run and the second run of each block, they were informed what the malfunction was and how they could respond. Pilots in the control group received all this information before the first run of each block, and they were informed that each subsequent run was a repetition.

The takeoff runs started with the aircraft on the runway (Runway 18C), with 3,000 ft of runway ahead. The U/V group was only instructed to respond as they saw fit, whereas the control group was told when the engine failure would occur and whether the takeoff was to be aborted or continued. Following the first run in a block, the U/V

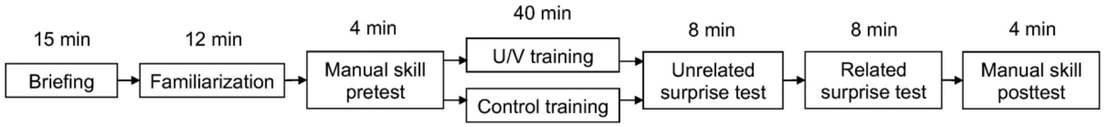

Figure 1. The experimental design. U/V = unpredictable/variable.

TABLE 3: Characteristics of the Runs

Run ID	Malfunction timing	Malfunction side	Wind direction from	Wind strength, kt
Engine failure during takeoff				
i	Gear lever up	Left	Left	9.7
ii[a]	Speed 65 kt	Right	Left	9.7
iii[a]	Rotate	Right	Left	9.7
iv	Altitude 270 ft	Right	Left	9.7
v	Gear halfway up	Left	Left	9.7
vi	Altitude 310 ft	Right	Left	9.7
Flyby runs with rudder failure				
i	20 s into run	15° right	Right	13.6
ii	50 s into run	20° right	Ahead	13.6
iii	50 s into run	25° right	Left	13.6
iv	30 s into run	10° right	Right	13.6
Flyby runs with engine failure				
i	20 s into run	Left	Left	9.7
ii	40 s into run	Right	Right	9.7
iii	30 s into run	Left	Left	9.7
iv	50 s into run	Right	Right	9.7

[a] In these runs, the takeoff was to be aborted.

group was reminded to pay attention to the minimum control speed (80 kt), below which the takeoff was to be aborted. If pilots continued the takeoff following the engine failure, they were instructed to continue to climb 100 ft straight ahead, where the run ended. Both groups included a similar number of to-be-aborted takeoffs.

The flyby runs started with the aircraft in approach, approximately 90 seconds from reaching the runway. The task was to fly over the runway, follow the centerline as closely as possible, and, upon reaching the runway, descend to 100-ft altitude and reduce speed to 85 kt. The gear remained down, and flaps remained at 25°. The malfunction occurred before reaching the runway. For the rudder issue, response instructions included that it could be countered by commanding a throttle differential. For the engine failure, pilots were reminded of the minimum control speed of 80 kt. In the U/V group, extra variation was introduced (besides factors listed in Table 3) by asking pilots to increase their speed in the second half of each flyby, by adding turbulence, and by reducing the visibility in 50% of the runs.

Related surprise test. After the practice session, two surprise tests were performed: first, an unrelated (control) surprise test, which is described in

TABLE 4: Order of the Runs and the Variations Used in the Practice Session

	U/V group		Control group	
Block: Run order	Scenario	Run ID	Scenario	Run ID
1				
1	Takeoff	i	Takeoff	i
2	Takeoff	ii	Takeoff	i
2				
3	FB rudder	i	Takeoff	i
4	FB rudder	ii	Takeoff	ii
3				
5	FB engine	i	Takeoff	ii
6	FB engine	ii	Takeoff	ii
4				
7	Takeoff	iii	FB rudder	i
8	Takeoff	iv	FB rudder	i
5				
9	FB rudder	iii	FB rudder	i
10	FB rudder	iv	FB rudder	i
6				
11	FB engine	iii	FB engine	i
12	FB engine	iv	FB engine	i
7				
13	Takeoff	v	FB engine	i
14	Takeoff	vi	FB engine	i

Note. FB = flyby; U/V = unpredictable and variable.

the next section; second, a related surprise test, which was the main test of the study. This test required the application of practiced skills (i.e., management of asymmetric thrust) in a surprising, demanding, and partly novel scenario. The scenario started on the runway of a different airport, featuring a single 4,000-ft-long runway (Runway 03) and a line of trees that was to be crossed following takeoff. There was moderate crosswind, coming from 310° (see Figure 2). Pilots were instructed to fly a left-handed circuit.

As shown in Figure 2, the following malfunctions were inserted into this run. First, during takeoff, when the speed reached 55 kt, thrust in the right engine dropped in 20 seconds to 40%. After a callout (or else after approximately 30 seconds), pilots were instructed to continue the circuit at 800 instead of 1,000 ft, to limit the run's duration. When 490 ft was reached, there was a brief dip in power (3 seconds) of the still fully functioning (left) engine. Pilots were immediately informed that both engines were unreliable and that they could keep using them both. This event was included to ensure that pilots were able to apply differential throttle as a means to stabilize the plane during the latter part of the run. Finally, the rudder's effectiveness decreased to 20% when pilots rolled out of the turn toward downwind, decreasing their ability to counter the thrust differential that was caused by the engine failure.

A successful landing was more likely if pilots identified the failures, compensated for the disturbances (potentially by commanding asymmetric throttle), and projected how the decrease in speed during the landing would affect these disturbances. Whereas the first two steps were practiced explicitly in the practice session,

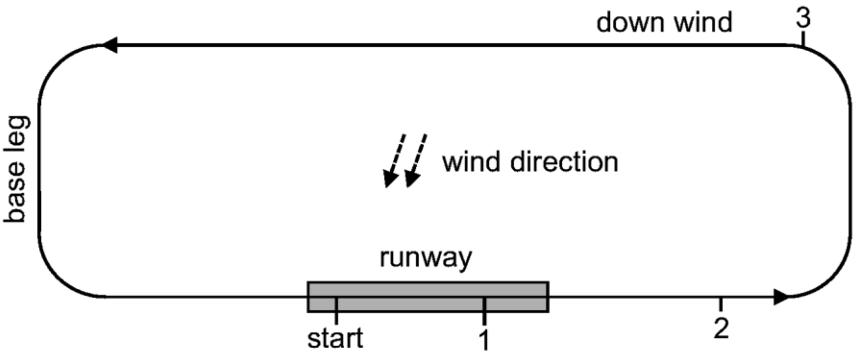

Figure 2. The circuit flown in the related surprise test and the moments at which the malfunctions occur. 1: Right engine starts losing power over 20 seconds. 2: Brief decrease in left engine power, which is immediately restored. 3: Rudder effectiveness decreases to 20%.

finding a solution for the landing was not. At the minimum control speed with a single engine (80 kt), the control surfaces can no longer be used to counter the moments resulting from the engine's asymmetric thrust. One solution is to maintain high speed during the landing, keeping the control surfaces effective. A second solution is to apply little throttle and make a steeper landing. As an additional measure, the thrust asymmetry caused by the engine failure can be reduced by commanding differential throttle, that is, by reducing power in the fully functioning engine.

Unrelated surprise test. An unrelated surprise test was included as a control test to evaluate whether the groups responded differently to a surprise event that did not feature the practiced principles. Although the groups were balanced as best as possible, inadvertent differences in preexisting skills (e.g., due to company training) or traits could still exist. The test was also included to provide both groups with a surprise so that potential differences in expectations were reduced between the groups. Finally, this test served to separate the related surprise test from the practice. The run started with the same instructions and settings as the related surprise test. From the moment of liftoff, the indicated airspeed decreased 1 kt/s from the actual airspeed. Pilots could identify the problem by checking whether the apparent decrease in speed corresponded with the pitch angle, vertical speed, control responsiveness, or ground speed. If pilots correctly called out the problem, they were instructed to finish the circuit. Otherwise, they were allowed to make an emergency landing.

Manual skill pre- and posttest. After the familiarization section, pilots performed a manual skill pretest, which was repeated as a posttest. This test was included to check whether manual flying skills in the groups were different at the start or whether they were differently affected by the practice. It contained a precision steering task requiring comparable manual skills as the related surprise test (i.e., performing a landing with ailerons only in moderate crosswind). The task started in approach, approximately 2 minutes from touchdown, with moderate crosswind from the left. It was announced that before the runway was reached, the rudder would become fixed in the neutral position. It was also announced that nose wheel steering would become inoperative due to this malfunction and that the nose should be pitched up as long as possible during the rollout. Pilots were asked to follow the glide slope (indicated by the PAPI lights or precision approach path indicators) and to land on the centerline as accurately as possible.

Dependent Variables and Hypotheses

Practice session. The time to correct callouts was obtained with the voice recorder. This time was defined as the time from the start of the malfunction to the utterance of the malfunctioning system (i.e., "engine," "rudder," or "speed indicator"). These times were measured in every first run of a block during the practice (see Table 4), as these runs were designed to be (maximally) different in predictability between the groups. The hypothesis was that the U/V group would have more difficulties with determining the problems, thereby causing longer correct callout times.

Further manipulation checks of the practice included the total time that participants were controlling the aircraft and the pilots' interest and enjoyment, as measured with the Interest and Enjoyment subscale (seven items) of the Intrinsic Motivation Inventory (Ryan, 1982). The outcomes of these checks should be similar for the two groups.

Tests. In the related surprise test, the main performance measure was whether pilots managed to successfully land on the runway. Second, the throttle setting of the fully functioning engine was measured when the minimum control speed of 80 kt was reached. This provided an indication of pilots' awareness of the problem of commanding too much throttle during landing. Finally, the proportion of time in which pilots applied differential throttle (in the effective direction) was computed in the final stage of the run (i.e., from the rudder failure to touchdown). A differential was defined as a left-right difference of at least 10% of the maximum throttle.

Correct callout times (see practice session) following the single-engine failure and the rudder failure were measured in the related surprise test and following the airspeed indicator malfunction in the unrelated surprise test. Incorrect or missing callouts were counted as missing cases.

In the manual skill pre- and posttest, the root mean square of elevator and aileron corrections was obtained in the last phase of the landing (35 seconds to 5 seconds before touchdown). To discard low-frequency components (e.g., caused by trim), these inputs were first high-pass filtered (forward and backward) with a second-order Butterworth filter with a cutoff frequency of 0.1 Hz.

At the end of the entire simulator session, pilots rated their experiences following the apparent airspeed problem (unrelated surprise test), the single-engine failure, and the rudder failure (related surprise test). From the latter two ratings, one general score was obtained by taking the mean of each pair. Subjective surprise and startle were rated on a 5-point Likert-type scale in answer to the following questions: "How surprised were you when you discovered the issue?" and "How startled or shocked were you when you discovered the issue?" (1 = *not at all*, 5 = *extremely*). Understanding was rated similarly by answering "How difficult was it to understand what had happened?" These scores were then reversed.

Improved performance, including faster correct callout times and higher reported understanding, was expected in the U/V group in the related surprise test. This group was expected to feel less threatened and less confused by the problems, causing them to report less startle (Martin, Murray, Bates, & Lee, 2015) and surprise (Foster & Keane, 2015). In the unrelated surprise test, the measures of performance were callout times and reported understanding. These measures, as well as reported surprise and startle, were not expected to differ between the groups in this test. The manual skill tests were expected to show an increase in manual skill from pre- to posttest due to familiarization with the controls. No other differences were expected, since the groups should be equally balanced and become equally familiarized with steering and landing due to the practice.

Data Analysis

Differences between the groups in the surprise tests were tested separately with independent-samples *t* tests or with Pearson's chi-square test in case of binominal data. Differences between the groups in correct callout times during the practice were tested with Group × Block mixed-model analyses of variance. Manual skill in the pre- and posttest was analyzed with Group × Test mixed model analyses of variance. Significant main effects of group and significant interaction effects were followed up with group comparisons. The significance level of reported significant results was set at $p < .05$. Holm-Bonferroni correction for multiple comparisons was applied separately to the performance measures, correct callout times, and subjective measures.

RESULTS

Manipulation Checks of the Practice

Callout times. Figure 3 shows boxplots of the correct callout times, and Table 5 lists the

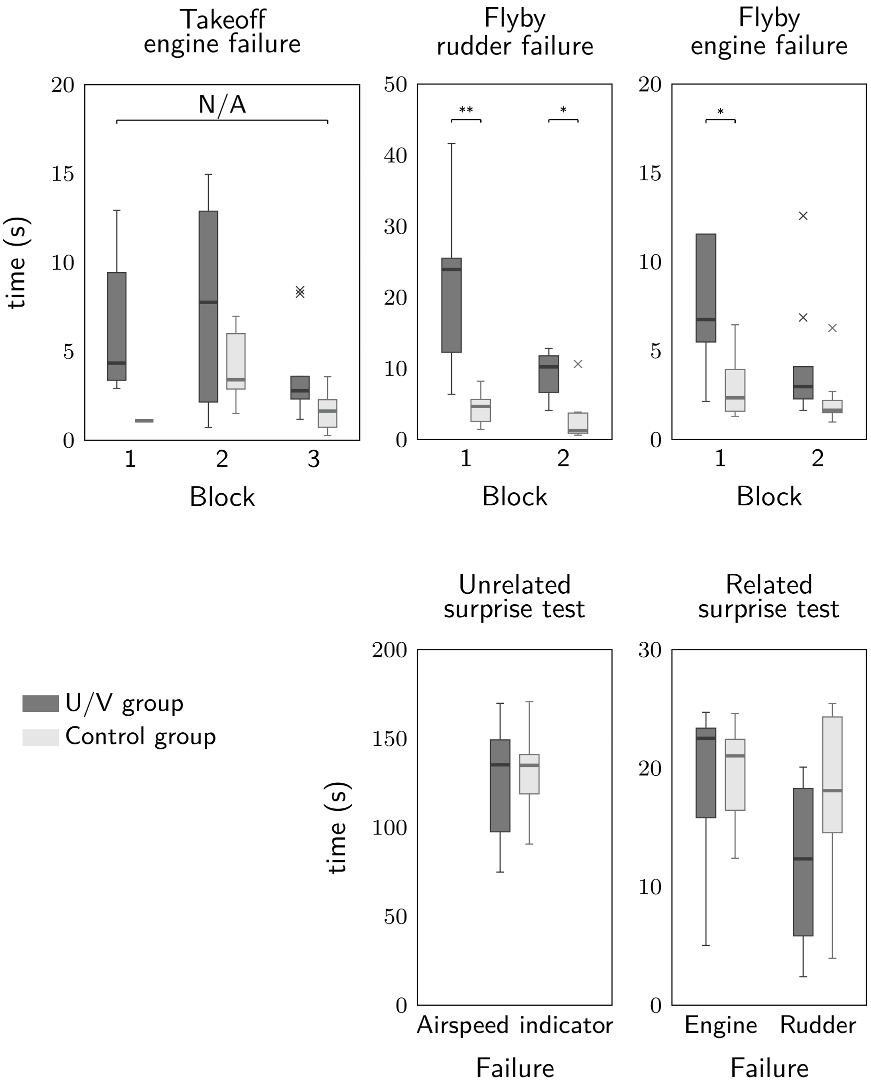

Figure 3. Tukey boxplots of the correct callout times in the practice session (top plots) and the unrelated surprise test and the related surprise test (bottom plots). Values are presented as median, interquartile range, and outliers. $^{*}p < .05$. $^{**}p < .01$.

corresponding statistical analyses. Correct callout times were overall longer in the U/V group than in the control group, indicating that the U/V group spent more time making sense of the events. In the flyby runs with rudder malfunction, this was the case in both blocks, whereas in the flyby runs with the engine failure, this was the case in the first block only. The takeoff runs with engine failure were excluded from statistical analysis due to an insufficient number of valid cases in the control group in the first run ($n = 1$). Missing cases in all scenarios resulted from pilots giving no callout, giving only an incorrect callout, or indicating that they did not know the cause of the problem.

Flight time. The time controlling the aircraft in the practice session did not differ significantly between the groups, $p = .670$. The mean flight time in the practice session was 29 minutes 47 seconds, $SD = 1$ minute 14 seconds.

Interest and enjoyment. . There was no significant difference in scores on the Interest and Enjoyment subscale of the Intrinsic Motivation Inventory, $t = .555$, $p = .586$, indicating that the control practice, mean = 43.90, $SD = 3.90$, was

TABLE 5: Statistical Analysis of the Correct Callout Times During the Practice

				Valid cases	
Failure	*F*	*df*	*p*	U/V group	Control group
Takeoff engine[a]	—	—	—	6	1
Flyby rudder (group)	12.86**	1, 10	.005	5	7
Flyby engine (Group x Block)	4.61*	1, 16	.047	9	9

[a]Excluded from analysis due to an insufficient number of valid cases in the control group.
$*p < .05$. $**p < .01$.

not perceived as less interesting than the U/V practice, mean = 44.70, *SD* = 2.36. Both groups rated the practice near the maximum score (i.e., 49), suggesting that the pilots generally found the practice interesting and enjoyable.

Related Surprise Test

Performance. The run was completed with a successful landing on the runway by 9 of 10 pilots in the U/V group and by 2 of 10 pilots in the control group. This difference was significant, $\chi^2(1, 19) = 9.90$, $p = .002$. One pilot in the U/V group and four pilots in the control group landed somewhere else. Four other pilots in the control group lost control in-flight, and the simulation was stopped due to an extreme attitude at a low altitude (<300 ft; see, e.g., Figure 4). Unsuccessful landings always involved moments of losing aileron authority. Three pilots in the control group responded to loss of aileron authority by increasing throttle, which exacerbated the problems. The average throttle setting when 80 kt was reached at the end of the circuit was significantly lower in the U/V group than in the control group (Table 6), meaning that the U/V group appeared to use a more favorable throttle/speed ratio during the landing. The U/V group also applied differential throttle during a larger proportion of the last part of the run (Table 6). Two pilots in both groups did not apply it at all, and two pilots in the control group applied it solely in the opposite direction.

Performance example. Figure 4 shows an example in which a pilot lost control. No differential throttle was applied during the run (middle plot). On base leg, flaps were set to 25°, and *gear down* was selected (top plot). This caused the speed to rapidly drop below 85 kt around 135 seconds (top plot). The pilot responded to this by increasing throttle (middle plot). In turning towards the runway, aileron authority was lost, as indicated by the increasing bank angle despite maximum inputs in the opposite direction (also positive per convention) at 140 seconds (bottom plot). Altitude was traded for speed at 145 seconds (top plot), and *gear up* was selected again to decrease drag. Despite these efforts, the decreasing speed caused loss of aileron authority again at 160 seconds and 170 seconds, after which the run was stopped to prevent a crash.

Callout times. There were no significant differences between the groups in correct callout times (see Table 6 and Figure 3). Missing values included four pilots who did not identify the malfunction and four pilots who did not give any callout, perhaps due to being too absorbed in the task. None of the missing cases involved incorrect callouts. When filling in the questionnaire after the test, all pilots indicated that they had noticed increasing controllability issues in downwind.

Subjective ratings. The events were rated as significantly easier to understand and as less surprising by the U/V group versus the control group (Table 6). Startle scores did not differ significantly between the groups, although there was a trend toward lower scores in the U/V group ($p = .063$). Pilots were on average moderately surprised (around 3.0) by the events, indicating that the surprise manipulation was successful. The maximum rating of surprise was 4 (very) for all events. Startle ratings were on average slight (2.0) to moderate (3.0). The maximum startle rating was 5 in the unrelated surprise test, 4 for the rudder failure, and 3 for the

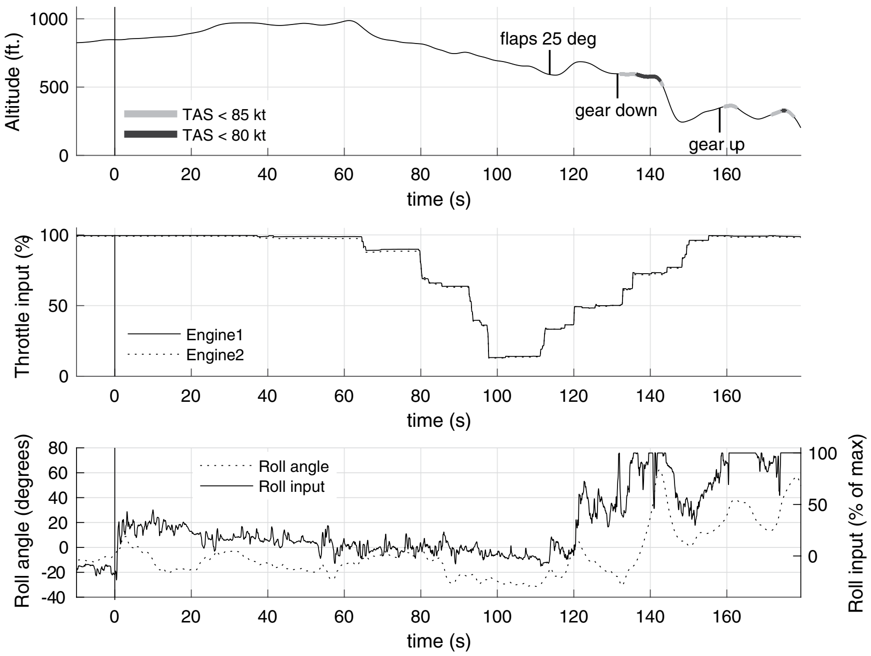

Figure 4. The altitude and speed (top), throttle input (middle), and roll angle and roll inputs (bottom) during a run with of loss of control in the related surprise test.

TABLE 6: Group Differences in the Related Surprise Test

	U/V group		Control group				
	Mean (*SD*)	*n*	Mean (*SD*)	*n*	Δ	*t*	*p*
Throttle at 80 kt, % (max)	19.05 (24.45)	10	56.12 (32.33)	10	37.07*	2.89	.010
Diff throttle applied, % (run)	45.62 (33.72)	10	18.08 (20.42)	10	–27.53*	2.21	.040
Correct callout time, s							
Engine failure	22.03 (9.81)	9	20.50 (5.14)	8	–1.52	0.39	.700
Rudder failure	10.88 (7.61)	5	24.07 (17.48)	7	13.19	1.57	.148
Surprise, 1–5	2.50 (0.74)	10	3.10 (0.39)	10	0.60*	2.57	.019
Startle, 1–5	2.10 (0.47)	10	2.65 (0.47)	10	0.55	1.98	.063
Understanding, 1–5	3.95 (0.69)	10	3.10 (0.66)	10	–0.85*	2.83	.011

*$p < .05$; results are significant after Holm-Bonferroni correction.

engine failure. One pilot in the U/V group broke off the takeoff. The scenario was repeated with the instruction to continue the takeoff, and the subjective ratings of the engine failure were obtained with regard to the first run.

Unrelated Surprise Test

Table 7 lists the outcomes of the unrelated surprise test. There were no significant differences between the groups in any of the measures. One pilot in the U/V group did not identify the problem and performed an emergency landing. One missing case in the control group was caused by a simulator malfunction during the run.

Manual skill pre- and posttest. In the posttest versus the pretest, both groups used significantly fewer inputs on the ailerons, $F(1, 18) = 7.29$, $p = .015$, and on the elevator, $F(1, 18) = 23.15$,

TABLE 7: Group Differences in the Unrelated Surprise Test

	U/V group		Control group				
	Mean (*SD*)	*n*	Mean (*SD*)	*n*	Δ	*t*	*p*
Correct callout time, s	118.46 (38.13)	9	136.36 (26.12)	9	17.90	1.16	.262
Surprise, 1–5	3.40 (0.84)	10	3.70 (0.48)	10	0.30	0.98	.345
Startle, 1–5	2.90 (1.10)	10	2.90 (1.10)	10	0.00	<0.01	>.999
Understanding, 1–5	3.00 (1.63)	10	3.30 (1.06)	10	0.30	0.49	.632

p < .001, indicating increased familiarization with the controls. There were no significant differences between the groups in aileron and elevator inputs (p = .522 and .354, respectively), nor were there significant Group × Test interaction effects (p = .421, p = .831, respectively), indicating that the practice had not affected manual skill of the groups differently.

DISCUSSION

The results of this simulator experiment show that pilots who had received the U/V practice used throttle and airspeed more effectively in a novel and unexpected situation, which resulted in more successful landings. The subjective ratings confirmed that the U/V group found it easier to understand the events in the test and reported, perhaps as a consequence, significantly less surprise (Foster & Keane, 2015). The control tests suggested that the results in the related surprise test were not attributable to preexisting differences between the groups or to differences in habituation to surprise or familiarization with the controls. Three pilots in the U/V group lost aileron authority during the practice session, which may have affected their performance in the related surprise test. However, the difference in successful landings between the groups is still statistically significant when these pilots are excluded. In addition, three pilots in the control group also experienced loss of aileron control early in the related surprise test (before turning to base leg).

The surprise ratings in the tests were generally higher than the startle ratings, indicating that the events were foremost unexpected but did not include very intense or threatening stimuli. Interestingly, the difference between the groups in startle and surprise ratings was similar in magnitude, but that of startle did not reach statistical significance due to a larger variance. Something similar was observed in a previous study (Landman et al., 2017b), so it may be indicative of a higher interpersonal variation in startle responses, or a larger variation in interpretation of the startle rating scale. In this respect, it is worth contemplating the value of averaging surprise and startle responses. Although it is required for the statistical analysis of training effectiveness, individualized pilot training may benefit more from the evaluation of individual surprise and startle responses.

A limitation of the study is that the practice session was very brief and the pilots were not trained to proficiency. There was little time between the practice and the surprise test, so the study provides no insight into long-term effects of U/V practice. Before U/V is implemented in pilot training, factors such as the optimal degree of U/V, which aspects of tasks to make unpredictable or variable, and the optimal stage of training to introduce U/V should be carefully considered. Finally, it cannot be ruled out that unintended differences between the groups existed.

The results are interesting in the light of contemporary theories on surprise and sensemaking (e.g., Klein, Moon, & Hoffman, 2006; Landman et al., 2017a; Zhang et al., 2008). According to this theoretical framework, the U/V group, as compared with the control group, was more stimulated to perform sensemaking activities during the training, which helped them to develop a better "frame" with regards to the effects of asymmetric thrust, control inputs, and airspeed on aircraft behavior. Since this frame was applicable in the related surprise test, it may have helped the U/V group to make sense of the events more quickly and extensively. In line with previous literature on frame or schema construction (e.g., Neisser, 1976; Schmidt, 1975), our results imply that obtaining knowledge about principles that overarch specific training experiences is essential for building resilient complex skills. The current study also suggests that U/V training is a means to obtain such resilience (see also, Van Merriënboer et al., 2002). Future research may be aimed at investigating whether more general problem-solving skills exist (e.g., "flexible procedures"; Field, Rankin, Mohrmann, Boland, & Woltjer, 2017), which can be effectively applied in situations that are entirely new and unpracticed.

In conclusion, the results show that organizing part of pilot training in a U/V way can be an effective means to improve the generalization of skills to in-flight situations that are not explicitly trained. Also, they suggest that one-sided and predictable training is insufficient as a means to prepare pilots for unexpected and novel situations.

ACKNOWLEDGMENTS

We thank Herman Koolstra, PhD, for helping us develop and tune startling flight scenarios. In addition, we are grateful for the useful comments provided by the reviewers to improve the final manuscript.

KEY POINTS

- The addition of unpredictability and variability to a simulator training session improved pilot responses in a surprise test that required an application of the practiced skills.
- Control tests suggested that the effects were not attributable to habituation to surprise or familiarization with the controls.
- The outcomes indicate that predictable and one-sided training scenarios are insufficient to prepare pilots for unexpected situations in-flight.

REFERENCES

Brady, F. (2004). Contextual interference: A meta-analytic study. *Perceptual and Motor Skills*, *99*, 116–126.

Bureau d'Enquêtes et d'Analyses pour la Sécurité de l'Aviation Civile. (2012). *Final report on the accident on 1st June 2009 to the Airbus A330-203, registered F-GZCP, operated by Air France, Flight AF 447 Rio de Janeiro-Paris*. Le Bourget, France: Author.

Caldwell, J. J. (1997). Fatigue in the aviation environment: An overview of the causes and effects as well as recommended countermeasures. *Aviation, Space, and Environmental Medicine*, *68*, 932–938.

Carbonell, K. B., Stalmeijer, R. E., Könings, K. D., Segers, M., & van Merriënboer, J. J. (2014). How experts deal with novel situations: A review of adaptive expertise. *Educational Research Review*, *12*, 14–29.

Casner, S. M., Geven, R. W., & Williams, K. T. (2013). The effectiveness of airline pilot training for abnormal events. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *55*, 477–485. doi:10.1177/0018720812466893

De Muynck, R., & Hesse, M. V. (1990, June). *The a priori simulator software package of the Piper PA34 Seneca III* (Technical report). Delft, Netherlands: TU Delft.

European Aviation Safety Agency. (2015). *Loss of control prevention and recovery training: Notice of proposed amendment 2015–13* (No. 120-111). Cologne, Germany: Author.

Eysenck, M. W., Derakshan, N., Santos, R., & Calvo, M. G. (2007). Anxiety and cognitive performance: Attentional control theory. *Emotion*, *7*, 336–353. doi:10.1037/1528-3542.7.2.336

Federal Aviation Administration. (1990). Advanced qualification program: Final rule. *Federal Register*, *55*(191).

Federal Aviation Administration. (2015). *Upset prevention and recovery training* (Advisory Circular No. 120-111). Washington, DC: Author.

Field, J., Rankin, A., Mohrmann, F., Boland, E., & Woltjer, R. (2017, June). Flexible procedures to deal with complex unexpected events in the cockpit. In *Proceedings of the Seventh REA Symposium, Sophia Antipolis Cedex, France*. Sophia Antipolis Cedex, France: Resilience Engineering Association.

Foster, M. I., & Keane, M. T. (2015). Why some surprises are more surprising than others: Surprise as a metacognitive sense of explanatory difficulty. *Cognitive Psychology*, *81*, 74–116. doi:10.1016/j.cogpsych.2015.08.004

International Civil Aviation Organization. (2013). *Manual of evidence-based training* (Document No. 9995). Montreal, Canada: Author.

Klein, G., Moon, B., & Hoffman, R. R. (2006). Making sense of sensemaking 2: A macrocognitive model. *IEEE Intelligent Systems*, *21*(5), 88–92. doi:10.1109/MIS.2006.100

Klein, G., Phillips, J. K., Rall, E. L., & Peluso, D. A. (2007). A data-frame theory of sensemaking. In R. R. Hoffman (Ed.), *Expertise out of context: Proceedings of the Sixth International Conference on Naturalistic Decision Making* (pp. 113–155). Mahwah, NJ: Erlbaum.

Koolstra, D. V. C. C., Herman, J., & Mulder, J. A. (2015). Effective model size for the prediction of the lateral control envelope of damaged aircraft. In AIAA Modeling and Simulation Technologies Conference (p. 2036). Reston, VA: American Institute of Aeronautics and Astronautics.

Landman, A., Groen, E. L., van Paassen, M. M., Bronkhorst, A. W., & Mulder, M. (2017a). Dealing with unexpected events on the flight deck: A conceptual model of startle and surprise. *Human Factors: The Journal of Human Factors and Ergonomics Society*, *59*, 1161–1172. doi:10.1177/0018720817723428

Landman, A., Groen, E. L., van Paassen, M. M., Bronkhorst, A. W., & Mulder, M. (2017b). The influence of surprise on upset recovery performance in airline pilots. *International Journal of Aerospace Psychology*, *27*(1–2), 2–14. doi:10.1080/10508414.2017.1365610

Lee, T. D., & Magill, R. A. (1985). Can forgetting facilitate skill acquisition? *Advances in Psychology*, *27*, 3–22.

Martin, W. L., Murray, P. S., Bates, P. R., & Lee, P. S. Y. (2015). Fear-potentiated startle: A review from an aviation perspective. *International Journal of Aviation Psychology*, *25*, 97–107. doi: 10.1080/10508414.2015.1128293

Neisser, U. (1976). *Cognition and reality: Principles and implications of cognitive psychology*. San Francisco, CA: Freeman.

Paas, F. G., & Van Merriënboer, J. J. (1994). Variability of worked examples and transfer of geometrical problem-solving skills: A cognitive-load approach. *Journal of Educational Psychology*, *86*, 122–133. doi:10.1037/0022-0663.86.1.122

Rankin, A., Woltjer, R., & Field, J. (2016). Sensemaking following surprise in the cockpit: A re-framing problem. *Cognition, Technology & Work*, *18*, 623–642. doi:10.1007/s10111-016-0390-2

Reid, L., & Nahon, M. (1988). Response of airline pilots to variations in flight simulator motion algorithms. *Journal of Aircraft*, *25*, 639–646. doi:10.2514/3.45635

Rohrer, D. (2009). The effects of spacing and mixing practice problems. *Journal for Research in Mathematics Education*, *40*, 4–17.

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. doi:10.1037/0022-3514.43.3.450

Schmidt, R. A. (1975). A schema theory of discrete motor skill learning. *Psychological Review*, *82*, 225–260. doi:10.1037/h0076770

Shapiro, D., & Schmidt, R. (1982). The schema theory: Recent evidence and developmental implications. In J. Kelso & J. Clark (Eds.), *The development of movement control and coordination* (pp. 113–150). New York, NY: Wiley.

Van Merriënboer, J. J., Clark, R. E., & De Croock, M. B. (2002). Blueprints for complex learning: The 4C/ID-model. *Educational Technology Research and Development*, *50*(2), 39–61. doi:10.1007/BF02504993

Zhang, P., Soergel, D., Klavans, J. L., & Oard, D. W. (2008). Extending sense-making models with ideas from cognition and learning theories. *Proceedings of the Association for Information Science and Technology*, *45*(1), 23–23. doi:10.1002/meet.2008.1450450219

Annemarie Landman received her MSc in Human Movement Sciences from the VU University Amsterdam in 2011. She is currently a PhD candidate at the faculty of Aerospace Engineering, TU Delft.

Peter van Oorschot received his MSc in Aerospace Engineering (Control and Operations) at the TU Delft in 2017.

M. M. (René) van Paassen received his MSc degree and PhD degree in Aerospace Engineering at the TU Delft, in 1988 and 1994, respectively. He is currently an associate professor at the Faculty of Aerospace Engineering, TU Delft.

Eric L. Groen earned his PhD in biology in 1997 from Utrecht University. He is currently is a senior scientist in the Multimodal Interface Research Group, Human Interfaces Department, at TNO Human Factors.

Adelbert W. Bronkhorst received his PhD in audiology in 1990 from the medical faculty of the VU University Amsterdam. He is currently a principal scientist at TNO Human Factors, and professor by special appointment of Applied Cognitive Psychology at the VU University Amsterdam.

Max Mulder received his MSc degree and PhD degree in Aerospace Engineering from TU Delft, the Netherlands, in 1992 and 1999, respectively. He is currently a full professor at the Faculty of Aerospace Engineering, TU Delft.

Date received: December 7, 2017
Date accepted: April 26, 2018