Pilot Trainees Benefit from Modelling and Adaptive Feedback

Yalmaz Ali Abdullah
EdTeKLA Group, Dept. of Computing Science
University of Alberta
Edmonton, Canada
yalmaz@ualberta.ca

Michael Guevarra
Illumia Labs
Winnepeg, Canada
michael.gue@illumialabs.ai

Minghao Cai
EdTeKLA Group, Dept. of Computing Science
University of Alberta
Edmonton, Canada
minghaocai@ualberta.ca

Jialiang Yan
Dept. of Computing Science
University of Alberta
Edmonton, Canada
jialian6@ualberta.ca

Matthew E. Taylor
Alberta Machine Intelligence Institute (Amii), Department of Computing Science
University of Alberta
Edmonton, Canada
matthew.e.taylor@ualberta.ca

Carrie Demmans Epp
EdTeKLA Group, Dept. of Computing Science
University of Alberta
Edmonton, Canada
demmanse@ualberta.ca

Abstract

Limited training capacity has contributed to a critical shortage of licensed commercial pilots. Adaptive educational technologies and simulators could alleviate current training bottlenecks if these technologies could assess trainee performance and provide appropriate feedback. Agents can be used to assess trainee performance, but there is insufficient guidance on how to provide concurrent feedback in simulation-based learning environments. So, we designed 4 feedback conditions that provide varying degrees of elaboration and used a within-subject study ($n = 20$) to compare feedback approaches. Trainee performance was best when they received highly-elaborative feedback that modeled expert behaviour. Variability in participant performance and preferences indicates a need to adapt the feedback type to individual learners and provides insight into the use of concurrent feedback in simulation-based learning environments. Specifically, learners appreciated the expert model because it facilitated a sense of control which was associated with lower negative affect and lower extraneous cognitive load.

CCS Concepts

• **Human-centered computing → User studies**; • **Applied computing → Interactive learning environments**.

Keywords

Feedback, Personalization, Schools/Educational Setting, Vocational Training, Simulators

ACM Reference Format:
Yalmaz Ali Abdullah, Michael Guevarra, Minghao Cai, Jialiang Yan, Matthew E. Taylor, and Carrie Demmans Epp. 2025. Pilot Trainees Benefit from Modelling and Adaptive Feedback. In *33rd ACM Conference on User Modeling, Adaptation and Personalization (UMAP '25), June 16–19, 2025, New York City, NY, USA.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3699682.3728337

1 Introduction

There is a critical shortage of pilots [56] which is amplified by an accompanying shortage of experienced pilots who can train new pilots. This is problematic given the forecast that we will be short nearly 80,000 pilots by 2032 [35]. In response to this shortage, some [33] have argued that adaptive educational technologies could supplement training. This technology-enhanced learning approach can employ an agent or other form of artificial intelligence (AI) to assess pilot trainee performance when using a flight simulator. Based on the agent's assessment, the system could then provide timely feedback to help trainees learn. However, we do not know how to best provision feedback in this context.

The provisioning of feedback in adaptive educational technologies has seen success in fields like mathematics [70], language learning [26], and computer science [5], where student learning has been linked to the sequencing of tasks [6, 46] as well as the scaffolding and feedback provided by the system [23, 63]. However, simulation-based educational technologies have received less attention. In particular, the study of how to provide visual feedback in simulator-based pilot training has been limited, which complicates the task of designing effective visual feedback [42, 63]. It also makes applying guidelines from other domains difficult [13]. We, therefore, examine four adaptive feedback approaches to provide insight into how they affect novice trainees' performance and experience.

Our results show that feedback augmented with signals that encourage learner modelling of expert behaviours better supported performance and was associated with lower levels of both negative affect and extraneous cognitive load. As we discuss later, this modelling-based feedback approach has the potential to support learners in similar environments. Additionally, our findings highlight the need for adaptive feedback tailored to individual learner orientations, prior knowledge, and self-regulation abilities.

UMAP '25, New York City, NY, USA

ACM ISBN 979-8-4007-1313-2/25/06
https://doi.org/10.1145/3699682.3728337

2 Feedback

In the context of pilot training, those with sufficient prior knowledge can use changes in their environment as a source of feedback. However, most novices lack this knowledge and would benefit from receiving formative feedback [63], which has two key features (1) verification, which provides correctness information, and (2) elaboration, which explains why a response is correct or incorrect. The effectiveness of formative feedback varies based on its timing [11, 43, 63]; the specificity and complexity of elaboration [63]; learner background-preparation [18] and motivation [63]; task features [43]; and instructional domain [68]. Our foci are timing, elaboration, specificity, and complexity since these are independent of individual learner traits.

2.1 Effect of Feedback Timing

Most work exploring feedback timing has focused on two options: (1) immediate, when feedback is given at the end of a task (e.g., a question), and (2) delayed, when trainees must wait to receive feedback (e.g., after completing a full test) [11, 43, 68]. This focus on immediate and delayed feedback may be due to the nature of the disciplines being taught (e.g., mathematics or English) and the materials that are used in most educational settings (e.g., books and paper tests).

Systematic reviews and meta-analyses have shown that the effects of these two timings differ [43, 68]. For example, immediate feedback was associated with faster learning in a grammar lesson, while delayed feedback was associated with better retention and performance on a delayed post-test [11]. This suggests that each timing might support different goals. In pilot-training contexts, delayed feedback is typically handled through debrief sessions with trainers [16], and immediate feedback is not extensively used because there is little time to pause for feedback between tasks during a flight exercise.

Going beyond this binary, simulators and flight exercises afford a third timing option: concurrent. Concurrent feedback is provided during a task and allows a trainee to change their response prior to task completion. This feedback could be ever-present (i.e., persistent), presented when a severe error is being made (i.e., bandwidth—exceeds a threshold [22]), presented at specified times (i.e., scheduled), or presented upon request (i.e., learner-controlled). Concurrent feedback in flight training is provided during a flight exercise where the trainer is present [16].

Prior work shows conflicting results when comparing concurrent and delayed feedback [49, 71]. Researchers have also reported differences when comparing across types of concurrent feedback [39]. Taken together, this indicates that the benefits of concurrent feedback are likely as nuanced as those for immediate and delayed feedback, suggesting a need for additional research in simulation-based training environments. Since simulators can be used to support replay and debrief sessions as part of supervised training, our investigation instead focuses on concurrent feedback for pilot trainees who are learning on their own.

2.2 Specificity and Complexity of Elaboration

Similar to timing [54], research on feedback elaboration has shown that the amount of elaboration may interact with learner traits, including their existing capabilities. High-ability students have been shown to perform better with pure verification feedback whereas low-ability students produced their highest scores when receiving more elaboration [34]. Motivation has also been linked to how learners respond to different degrees of elaboration [21]. Students with high performance-oriented learning goals (i.e., a desire to be positively evaluated by others) were shown to perform better with more elaborate feedback in a decision-making simulation [21].

Specificity and complexity are sub-features of elaboration. Feedback that lacks specificity can leave students uncertain about how to respond, increase cognitive burden, cause frustration, and reduce motivation [63, 73]. But feedback that is too specific or presented poorly can increase cognitive processing demands [63]. Research on feedback complexity in academic subjects has produced similarly inconsistent results: some studies found no effect [64] and others found negative effects [48]. Very little of this work has examined the effect of feedback specificity and complexity with respect to concurrent feedback. Work, such as that done by Davis et al. [21], only examines immediate feedback, which limits our understanding of how individual trainees will respond to different degrees of specificity and complexity in an environment that places a higher cognitive load on learners, as is the case in flight simulators.

3 Adaptive Pilot Training System

Reinforcement-learning (RL) agents have been used within adaptive training systems for a range of tasks including student skill development [32, 53, 61] and improving teaching strategies [72]. Guevarra et al. [33] detail an approach to using RL agents for providing formative feedback to pilot trainees where an agent that flies the aircraft [2, 55, 60] serves as an expert model to enable the identification of learner errors. Building on this idea, we created an adaptive training system that consists of three main components: (1) a flight simulator; (2) an RL agent that evaluates student performance; and (3) visual feedback to support learner self-correction.

We used the X-Plane 11 simulator for the first component because high-fidelity simulators can support learning [18]. This desktop simulator allows users to control the pitch and roll of an aircraft as illustrated in Figure 1 and 2.

Component 2 integrates a behavioural cloning agent as the tutor's pedagogical agent [7]. In our case, the agent was trained to mimic a flight instructor using imitation learning with Stable-Baselines3 [36, 59]. A recording of the flight instructor performing a fundamental flight task called "Straight and level" (see Section 3.1) served as the basis for the expert model. Following training, the pedagogical agent observes the environment as pilot trainees interact with the simulator, comparing trainee actions with the actions the agent expects to see based on its representation of expert performance. That assessment is then used to drive some forms of concurrent feedback.

Component 3 is the primary focus of the present study. It enables different approaches to providing feedback about trainee performance. These approaches are described in Section 3.2.

3.1 Learning Task

Straight and level (SnL) is an early maneuver that pilot trainees are taught. As the name implies, it consists of keeping the aircraft level

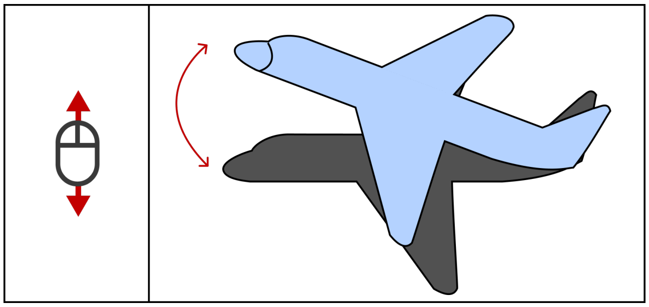


Figure 1: Rotation around the pitch axis can be adjusted by moving the mouse vertically.

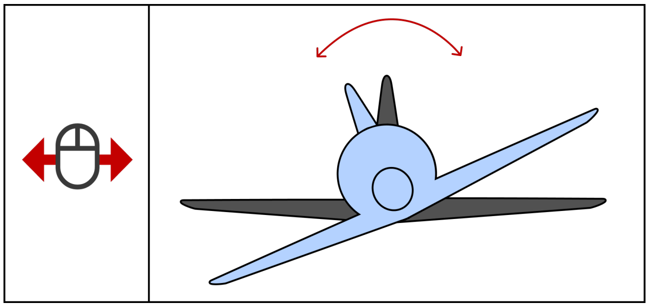


Figure 2: Rotation around the roll axis can be adjusted by moving the mouse horizontally

Table 1: Feedback attributes (Verif. - Verification; Elab. - Elaboration; Specific. - Specificity; Complex. - Complexity)

	Verif.	Timing	Elab.	Specific.	Complex.
B	No	None	None	None	None
BG	Yes	Bandwidth	Low	Low	High
DG	Yes	Bandwidth	Moderate	Moderate	Moderate
DD	Yes	Persistent	Very High	High	Low
TA	Yes	Persistent	High	High	Low

while flying straight towards a target. Advanced tasks build upon the foundational skills learned from SnL.

Studying trainee responses to feedback for this fundamental task allows us to better control for factors such as model accuracy and task difficulty; laying the groundwork for complex scenarios. The simplicity of the task also allows us to study how novice trainees (our intended learners) might respond.

3.2 Feedback Options

All 5 feedback conditions are described below and summarized in Table 1. These conditions were designed by a team of pilots, pilot instructors, and researchers based on Shute's guidelines [63]. For this study, we chose to focus on visual feedback because it aligns more closely with the training practices already used in flight-training environments, where a system like ours would be integrated into the learning process.

3.2.1 B: baseline (control): There is no verification, diagnosis of trainee errors, or provisioning of explicit feedback. This condition mirrors current simulator practices, where only implicit feedback is provided through the simulated environment and physics system.

In this case, the world is updated to reflect trainee actions and the trainee must know how to interpret the world and respond in an appropriate way. All other feedback conditions provide verification to trainees.

3.2.2 BG: binary glow (Figure 3a): This feedback colours the edges of the screen yellow when the pitch or roll of the aircraft exceeds acceptable thresholds as determined through comparison to the agent. The strength of colouring is proportional to the degree of deviation. If the trainee corrects their mistake, the glow goes away.

Error thresholds for both this and the directional glow condition (see below) are based on regulator-defined error tolerances for the flight task [17].

3.2.3 DG: directional glow (Figure 3b): This feedback is similar to binary glow but glow is only added in the direction of deviation. Thus, it suggests the direction of correction; the trainee must determine magnitude.

3.2.4 DD: dancing dots (Figure 3c): This feedback is displayed in a black square near the center of the screen. There are two coloured lines drawn on this square: the green line represents trainee input and the teal line shows the expert model for how to respond to the current situation.

Trainees should be able to use this representation to identify the difference between what they are doing, what they should be doing, and adjust their actions accordingly. This approach is similar to mixed-reality techniques for motor-skill training [28] that model expert performance through a visual representation (e.g., a video overlay of a silhouette), which has helped improve learner motor-skills in sports-training contexts [44].

3.2.5 TA: target alignment (Figure 3d): This feedback option is based on current flight instruction guidelines [1]. It provides additional visual support for something pilot-trainees are supposed to do. As such, it is an augmentation of the baseline that does not rely on the agent.

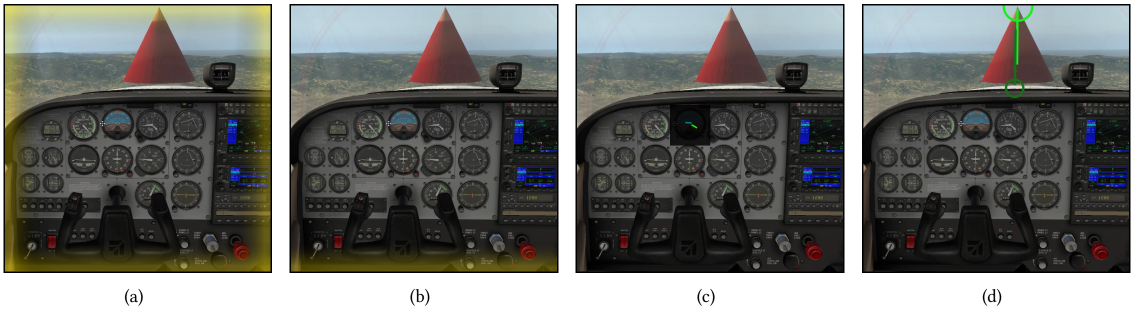
This cognitive support marks a reference point on the target and the nose of the aircraft using green circles. The lines are meant to help align these points vertically while maintaining a fixed distance between them. Trainees should be able to use the markers to identify corrective actions.

4 Methods

This mixed-methods, within-subject study counterbalanced trainee exposure to the feedback conditions and collected information about trainee performance, experiences, and preferences. All participants who consented and began a study session received $30.

4.1 Study Procedures

After giving consent, participants received instruction on how to control the aircraft and perform the flight task. Because one's affective response can impact learning [14], attention, and decision making [52], a baseline measurement of trainee affect was taken.


(a)

(b)

(c)

(d)

Figure 3: Feedback conditions (a - BG; b - DG; c - DD; d - TA)

We call this incoming (I). The measure was taken using the short form of the international positive and negative affect schedule (I-PANAS-SF) [66]. It consists of two subscales, positive affect (PA; α = .75) and negative affect (NA; α = .76), which are measured through a 10-item questionnaire and rated using a 5-point Likert scale (1-Strongly disagree to 5-Strongly agree). Scores for each subscale range from 5 to 25.

Following this, participants were exposed to each feedback condition, within which they attempted SnL three times. After completing the flight tasks within each condition, their affect was measured using the I-PANAS-SF. We also measured extraneous load [65] at this point to understand how feedback features interact with the already high cognitive demands of flight tasks [31]. We used Leppink's questionnaire [47], adapted to our domain, which has two Likert-scale items dedicated to measuring extraneous load. These were rated from Strongly Disagree (1) to Strongly Agree (10) and the average of both items was taken as the score. Because cognitive load is constrained by working memory capacity, this participant attribute was measured at the start of the study using the Corsi block tapping task [41]. The score for this can range from 2 to 9.

After all conditions were completed, participants were asked to rank the feedback conditions in order of perceived benefit and explain their rankings. Demographic information was then collected, which included participant age, gender, experience with games, experience with flight simulators, and experience with physical planes.

For counter balancing we used a Latin-square generator [62]. Given expected effect sizes, we recruited enough participants to complete two balanced Latin squares (n = 20). Participant data was reviewed immediately after collection. If any issue was detected, we repeated that sequence in the Latin square with another participant until we had one valid sample for each sequence in the squares.

A licensed pilot trainer scored flight performance on a 4-point scale [17], where 4 is the best score possible. We summed scores across attempts to obtain the condition score.

Study instruments are available via OSF[1]

4.2 Participants

Following institutional approval, we recruited 24 participants through a mailing list. Of these, 4 were excluded because they intentionally crashed the plane or otherwise did not follow instructions, an error in the simulator caused the application to crash, or the experimenter made an error when assigning the condition sequence.

Data from the remaining 20 were analyzed. Their ages ranged from 20 through 38 years (M = 23.2, SD = 3.76). Their working memory capacity ranged from 4 through 9 (M = 6.5, SD = 1.32), which is within the range expected for healthy adults [41, 69].

4.3 Analysis Procedures

As is common with mixed-methods approaches [20, 45], we complement quantitative analyses of trainee performance, preferences, and experience with qualitative analyses of information about trainee preferences. This triangulation accounts for the methodological weaknesses of each approach [67].

We used Shapiro-Wilk to check the normality assumption. When violations were detected, the Q-Q plots were examined to see whether these violations were within the tolerances of the statistical tests we planned to use (e.g., ANOVA). Highly-skewed distributions were corrected with $log(x+1)$ transformation [58]. When this transformation allowed the normality assumption to hold, we proceeded with parametric statistical tests. When it did not, we applied the equivalent non-parametric test to the untransformed data. In all cases, we report descriptive statistics as violin plots of the untransformed data to facilitate interpretation.

We used the rstatix library [40] to conduct tests. One-way repeated measures ANOVAs were used to test for differences across conditions. We report η^2 as a measure of effect size and applied Greenhouse-Geisser correction when Mauchly's test revealed a violation of the sphericity assumption. When data did not meet the normality assumption, we instead used the Friedman test and report Kendall's W as the effect size. Significant results ($alpha$ = .05) were followed by pairwise comparisons using paired t-tests for ANOVA and Wilcoxon Signed Rank tests for Friedman. We report r or Cohen's d as effect sizes and controlled for multiple comparisons using Holm's method [37] .

To analyze preference rankings, we performed a Chi-square test to examine whether the pattern of ranks was random (as recommended by Finch [29]), provided a rank frequency chart for each condition, and provided a pairwise rank comparison table. We also provided Dowdall scores [30] which are calculated by taking the sum of the reciprocal for each rank given to a condition. In our case, scores ranged from 1 to 0.2, with 1 indicating a higher rank.

[1] https://osf.io/2cjv9/?view_only=b7cd5d32edcd439bb059525ae18575f3

Explanations of preference rankings were analyzed using thematic analysis as described by Braun et al. [9]. The first and last authors familiarized themselves with the data before deciding on codes using a consensus-based approach. They then independently coded the data and reconvened to identify themes (Section 5.5).

Our analysis scripts and codebook have also been included in our OSF project.

5 Results

We found differences in trainee performance, the cognitive demands and affective responses associated with feedback conditions, and trainee preferences.

5.1 Differences in Performance

The Friedman test detected a small difference ($\chi^2(4) = 11.8, p = .019, W = .15$) in trainee performance across conditions. The subsequent Wilcoxon signed-rank tests (Table 2) indicated trainees performed better when in dancing dots than they did in some of the other conditions.

Table 2: Differences in trainee performance between feedback conditions (*p* adj = corrected *p* value)

Group 1	Group 2	*W*	*p*	*p* adj	*r*
B	BG	65.50	.916	1.000	0.06
B	DD	2.50	.006	.057	0.63
B	DG	84.00	.176	.880	0.35
B	TA	92.50	.769	1.000	0.08
BG	DD	21.00	.049	.346	0.47
BG	DG	72.50	.060	.357	0.44
BG	TA	88.00	.597	1.000	0.16
DD	**DG**	**153.00**	**.003**	**.032**	**0.65**
DD	TA	110.50	.026	.212	0.50
DG	TA	77.50	.738	1.000	0.05

Note the maximum possible score was not achieved in the baseline condition but was in other conditions (See Figure 4), suggesting that each of the experimental feedback conditions provided value to some trainees.

5.2 Differences in Extraneous Load

There were large differences across feedback conditions for extraneous load: $F(4, 76) = 7.80, p < .001, \eta^2 = .22$. We see considerable variability in the distributions for this measure (Figure 5), with dancing dots imposing the least extraneous load (Table 3). The extraneous load distribution for dancing dots appears to be bimodal, which indicates a group of participants experienced levels of extraneous load that were similarly high to those seen in other conditions.

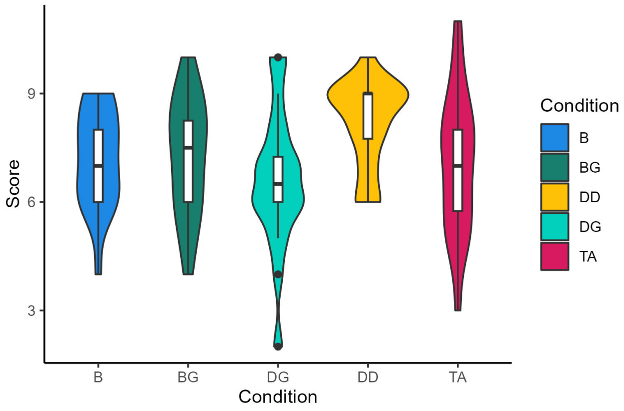


Figure 4: Performance score distributions by condition.

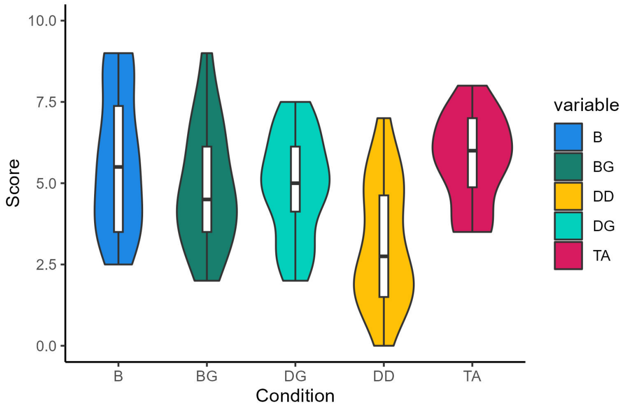


Figure 5: Extraneous load distribution by condition.

Table 3: Differences in extraneous load between conditions

Group 1	Group 2	*t*	df	*p*	*p* adj	*d*
B	BG	1.74	19	.099	.494	0.39
B	**DD**	**4.60**	**19**	**< .001**	**.002**	**1.03**
B	DG	1.53	19	.143	.500	0.34
B	TA	-0.04	19	.968	1.000	-0.09
BG	DD	2.94	19	.008	.059	0.66
BG	DG	-0.06	19	.957	1.000	-0.01
BG	TA	-1.60	19	.125	.500	-0.36
DD	**DG**	**-3.20**	**19**	**.005**	**.038**	**-0.72**
DD	**TA**	**-4.51**	**19**	**< .001**	**.002**	**-1.01**
DG	TA	-1.84	19	.082	.493	-0.41

5.3 Differences in Affective Experiences

No measurable difference was detected across conditions for positive affect ($F(2.96, 56.32) = 1.07, p = .37, \eta^2 = .01$), but differences were detected for negative affect ($F(3.49, 66.22) = 6.31, p < .001, \eta^2 = .09$). When considered alongside the descriptive statistics for positive affect (Figure 6) and negative affect (Figure 7), these ANOVA results suggest the selection of feedback only risked increased negative affect with little potential benefit to positive affect.

Pairwise comparisons of participants' negative affect across conditions (Table 4) show dancing dots had the lowest observed negative affect ($M = 2.2$, $SD = 2.64$) and the baseline condition was associated with high negative affect ($M = 4.8$, $SD = 3.01$).

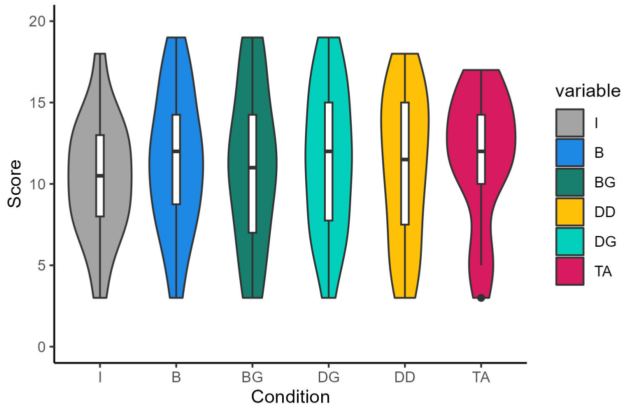


Figure 6: Positive affect distribution by condition.

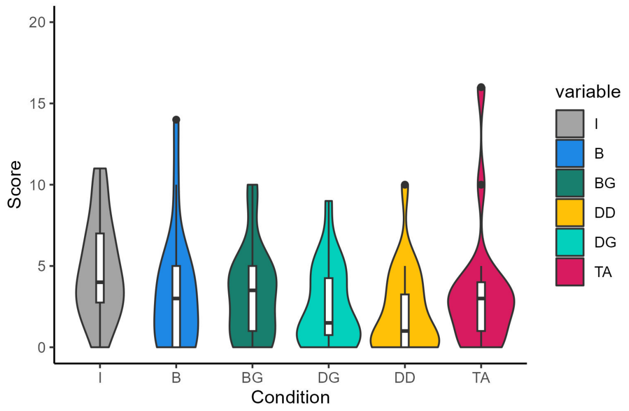


Figure 7: Negative affect distribution by condition.

Table 4: Differences in negative affect between conditions

Group 1	Group 2	t	df	p	$p.adj$	d
B	BG	-0.49	19	.628	1.000	-0.11
B	DD	1.66	19	.113	.904	0.37
B	DG	0.93	19	.363	1.000	0.21
B	I	-3.27	19	.004	.052	-0.73
B	TA	-0.54	19	.593	1.000	-0.12
BG	DD	2.88	19	.010	.106	0.64
BG	DG	1.34	19	.196	.980	0.30
BG	I	-3.11	19	.006	.070	-0.70
BG	TA	0.17	19	.867	1.000	0.04
DD	DG	-1.48	19	.156	.959	-0.33
DD	**I**	**-4.20**	**19**	**< .001**	**.007**	**-0.94**
DD	TA	-2.31	19	.032	.292	-0.52
DG	**I**	**-3.95**	**19**	**<.001**	**.012**	**-0.88**
DG	TA	-1.55	19	.137	.959	-0.35
I	TA	2.76	19	.013	.125	0.62

5.4 Differences in Trainee Preferences

The Chi-square test indicated there was a non-random pattern in the ranks given to each feedback condition, $\chi^2(4, N = 20) = 36.92$, $p < .001$. Dowdall scores, provided in Table 5, show dancing dots was most preferred and baseline was least preferred, suggesting that participants appreciated having some feedback during the task.

Table 5: Dowdall score by condition

	B	BG	DG	DD	TA
Dowdall Score	4.73	6.88	8.98	16.58	8.48

Looking at the rank frequency chart (Figure 8), which shows how often a condition was assigned each rank, confirms this interpretation. Baseline was never ranked first and dancing dots was ranked first by nearly 70% of participants, making it the highest rated condition by rank counts.

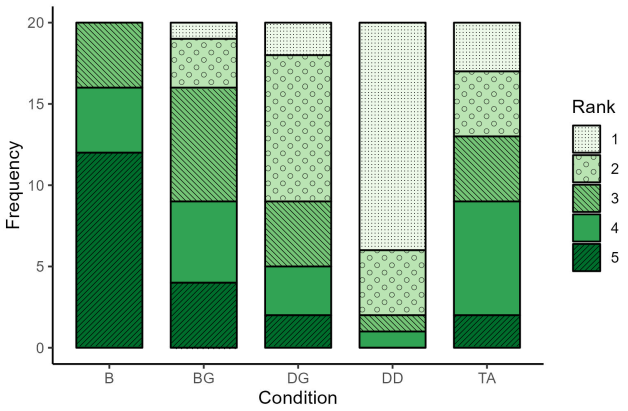


Figure 8: Rank frequency for each condition

Table 6 shows the frequency with which participants ranked one condition over another. Each cell in the table reflects how often a given condition (the row label) was ranked higher than another condition (the column label). It shows that dancing dots was ranked above the baseline condition in all cases and target alignment was ranked above the baseline in 17 cases. Both the dancing dots and target alignment conditions provided extensive elaboration. As such, these rankings indicate participants had a preference for elaborative feedback.

Table 6: Pairwise rank comparison

	B	BG	DG	DD	TA
B	-	5	4	0	3
BG	15	-	6	2	9
DG	16	14	-	3	13
DD	20	18	17	-	16
TA	17	11	7	4	-

5.5 How do users experience and respond to the feedback conditions?

Participant explanations of their feedback rankings fell into five interacting themes. They commented on how the specificity and nature of elaboration contributed to *feedback constructiveness* and how the evaluated feedback types impacted their *workload*, *sense of control*, and *perceived performance*. These, in turn, related to how they responded to the feedback through their behaivours and their associated *affective responses*.

5.5.1 Perceived Performance. Participants distinguished between their performance and how much they seemed to learn. As P1 explained, "the other feedback styles were helpful in allowing me to reach the goal, but I don't think they improved my ability to complete the task without them". These perceptions played a role in participant preferences, as they appeared to prioritize either their perceived learning or performance.

When prioritizing performance, participants chose the feedback condition that made it "easiest to stay level" (P11), "felt the easiest" (P3), or where they "did the best" (P3). When prioritizing learning, they chose "the most easy ones to follow ... and [the ones that] helped [them] understand how to fly the plane" (P5), with P20 commenting on how they had learned to fly from the feedback they received prior to the final condition. P20's final condition was the baseline that provided no additional feedback.

5.5.2 Sense of Control and Feedback Responsiveness. Participants said their ability to control the plane influenced their preferences. Participants disliked feedback that felt unresponsive (e.g., "glows were ... not the quickest when it came to input", P13). They also disliked feedback that made them feel like they had less control, as was the case for target alignment because it "was easy to overshoot" (P2) and they could not "understand how much adjustment was needed" (P19). In contrast, participants reported feeling "more in control" (P15) using dancing dots because it "resulted in the most precise and accurate results" (P19). This aligns with our analysis of preference rankings, which identified dancing dots as the most preferred feedback condition.

5.5.3 Workload. The cognitive workload imposed by a condition appeared to be closely tied to participants' perceived performance and sense of control. Participants expressed a preference for feedback that supported "efficiency in terms of how easily you could tell you were succeeding" (P7) and allowed them to "figure out the sensitivity of [the] controls" (P6). Such feedback reduced their workload and facilitated its practical use.

In general, simple feedback helped trainees focus on adjusting their performance by directing their attention to key aspects of the task, such as keeping the plane level (directional glow; P4) or flying "the plane slowly and smoothly" (dancing dots; P5). A lack of "visual clutter" (P20) and fewer "distractors" (P14) further supported their ability to concentrate on the intended task. Trainees favoured feedback that "felt the easiest" (P3) or was "straightforward" (P9). Conversely, conditions were liked less when they were "distracting and made it hard to tell what [trainees] needed to adjust" (P20), caused them to "shift back and forth" (P14), or "required more overthinking which, in turn, led to worse control of the plane" (P7).

These findings align with the preference rankings and cognitive load results, which showed that dancing dots imposed less cognitive load than most conditions and was the most preferred.

However, reduced workload was not always beneficial. Some noted that easier-to-use feedback allowed them to focus by "ignoring everything else" (P4). While this might support learning within a simulator, it could pose a high risk in real-world scenarios. Notably, participants who prioritized learning over performance viewed this as a drawback, stating that "directional glow is a little more difficult [to use], but engaged [them] with flying more" (P2).

5.5.4 Affective Responses. Similar to workload, participants' affective responses seemed to be driven by whether they could use the provided information to make sense of what was happening to adjust their actions in a way that supported their performance. In general, higher-workload feedback conditions, those that provided less control, and those that participants felt provided insufficient information were associated with less productive affective states. This includes feedback that made the trainee "hyperaware" (target alignment; P10), was "irritating" (BG; P12), was "frustrating" (binary glow, P19); or was "confusing" (directional glow; P8). These responses came from feedback that was considered "vague" (P11), feedback that "said you were doing it wrong and did not help you to fix [it]" (P12), or feedback that was "difficult to understand" (P18) from the perspective of what correction was needed.

Few purely positive affective responses were reported. Positive responses were typically tied to reductions in workload: "I felt you weren't distracted by the glow on the sides and I feel it made me less anxious" (P10). This reduction in negative affect was used to justify the provided rankings, with feedback types that reduced negative affect being preferred to those that were associated with increased negative affect or confusion.

5.5.5 Feedback Constructiveness. Across conditions, participants noted that the amount of information (elaboration) and specificity in feedback influenced their ability to use it effectively and make corrections (i.e., how constructive they perceived the feedback to be). In general, feedback that provided sufficient elaboration and specificity supported trainee interpretation, decision-making, and action-taking; it was, therefore, perceived as more constructive. Participants appreciated feedback that "guid[ed them] if [they] made a mistake so that [they] could better understand what a correct alignment would look like" (P1). For example, directional glow helped participants "understand [what] direction [they] should stop going in" (P5), while dancing dots "told you exactly what you needed to do to fly the plane straight" (P9). Conversely, less specific feedback (e.g., binary glow) elicited frustration, with comments like, "it failed to provide information about the direction I had to go to fix it" (P2).

Trainees' prioritization of learning versus performance, as described in Section 5.5.1, also shaped their perceptions of feedback elaboration and specificity. Feedback that indicated when they were "doing something wrong but still allow[ed them] to use the nose of the plane as a guide" (P1) improved self-efficacy because they felt the more elaborate and specific feedback hindered their ability to complete the task independently, with P1 remarking, "I don't think they improved my ability to complete the task without them".

These examples highlight a desire for different levels of information, with participants differing on whether the same feedback condition provided or lacked the necessary details to adjust their behaviour. This variation in perceived constructiveness was further influenced by the visibility and timeliness of the feedback. Some feedback types were disliked because they were "not the quickest when it came to input" (P13), while others were overly responsive, leading to participants "jerking the controls randomly" (P2). For glow-based conditions, the gradual fading in and out of the feedback appeared to make small corrections less noticeable. As one participant explained, "it [was] not specific... I also was struggling to notice the light flashing" (P16).

One factor that consistently impacted perceived constructiveness was the presence of verification. Participants reported that the only condition with no verification (the baseline) was "the least useful since there were no indicators at all" (P9). This aligns with our analysis of both preference and performance data, which demonstrated that having some form of feedback was always preferred in addition to the provided feedback supporting task performance.

6 Discussion

Most examinations of feedback in adaptive training systems focus on immediate or delayed feedback. Ours instead explored the ability of simulator-based training environments to provide different types of concurrent feedback and examined their impact on trainee performance, preference, cognitive load, and affect. We found that dancing dots (an expert modelling approach) supported better performance while also resulting in lower (extraneous) cognitive load and negative affect. Moreover, 70% of participants ranked this feedback condition as their preferred approach and participant reports indicate they preferred feedback that either improved their perceived performance or perceived learning.

6.1 Learning and Performance

Jointly, participant open-ended responses and task performance indicate that they learned the desired skills within the 30 minute training session. This rapid skill acquisition aligns with findings from other situated-learning settings [25] and a meta-analysis of simulator use in higher-education [18]. It is suggestive of the potential for retention and transfer to other flight tasks or real-world flight settings [12, 19, 25]. Whether trainees retain the skills they exhibited or can apply them in real-world flight tasks remains to be seen.

Examination of participant performance across feedback conditions shows that most feedback designs neither helped nor harmed trainee performance, in the short term. This could be related to differences in individual trainee backgrounds and habits or feedback design, with all of these factors being known to interact with learning and learner behavioural patterns [3, 15, 27].

6.2 Elaboration and Modelling Support Trainees

Trainees' stronger performance in the dancing dots condition provides insight into feedback design for situated learning for skill development. It highlights the value of learning through modeling expert behaviour, aligning with findings from other contexts [4].

Dancing dots gave the pedagogical agent a privileged status that made it desirable for trainees to try and model its behaviour. The interface design, which presented a simplified and uniform representation (the two dots) of trainee and expert behaviour, was able to provide elaborated feedback and encourage the trainee to engage in behavioural modelling. Trainee comments reflect this, noting that dancing dots gave them the information they needed to adjust their behaviours to match the desired behaviours. Moreover, the benefit of combining this elaboration with modelling is reflected in their poorer performance in the target alignment condition, where they received similar amounts of information via an alternative representation but were not shown a behavioural model.

Our findings suggest the potential for improving human learning by augmenting feedback in simulators with signals that encourage learner modelling of expert behaviours [15]. Specifically, the learning of contextualized, dynamic, and complex tasks in other simulation environments, where appropriate learner responses similarly depend on continually updating contextual variables [42], could be better supported through this modelling feedback approach.

6.3 Experiencing Feedback

Examining how trainees attended to information provided by different feedback conditions revealed that they often ignored aspects of feedback to focus on other dimensions. A similar behaviour was observed in studies of delayed feedback [26], where this learner strategy compensated for AI errors. However, in our study, it emerged as a maladaptive self-regulation strategy that trainees employed when overwhelmed by the task. In the present study, the use of this strategy interfered with learning and was characterized by trainees alternating between pitch or roll correction. These findings underscore the need to reconsider how feedback is designed and provisioned in these higher-demand learning environments.

6.4 Feedback Provisioning Needs to be Adaptive

While the modelling feedback (i.e., dancing dots) was superior to the other forms of feedback, the results across conditions for trainee performance, cognitive load, affect, and experience indicate a need for personalization and adaption in feedback provisioning. For example, trainee feedback preferences aligned with their learning goals: those with performance-oriented goals favoured elaborative feedback (e.g., dancing dots), while those with learning-oriented goals (i.e., the desire to develop new skills and competencies) preferred simpler designs (e.g., directional glow) [21]. By adjusting the type of feedback to align with learner orientation, we could help support learner motivation [38] and potentially scaffold the transition from a performance to a learning orientation.

However, identifying the optimal timing and type of feedback for this sort of adaptivity requires further research. In the interim, adaptation can be left in the hands of the trainee. Giving them control over which feedback to use could support adoption by meeting their affective needs [57], but trainees' limited self-regulatory skills [74] and knowledge [11, 51, 51] may require additional guidance through features, like open-learner models and learning dashboards [8, 10, 23, 24, 50, 51]. Trainer-controlled feedback is another option, where trainers could configure the concurrent feedback settings to help pilot-trainees progress. Data collected from the above settings could then be used to understand learner responses to feedback and build an agent for adaptively provisioning appropriate feedback.

Based on a combination of the literature and our findings, we suggest going beyond selecting the feedback approach to changing the learning task based on each trainee's prior knowledge. The bimodal distribution of extraneous cognitive load observed in the dancing dots condition indicates that we had two groups of trainees who varied in some combination of their cognitive abilities and prior knowledge. Since it is well established that completing tasks better supports learning for those with sufficient background knowledge, [65], we could start with worked examples of flight tasks for those with less prior knowledge and increase the level of detail provided in the simulation for those with more background preparation [18].

7 Limitations

This study used counter-balancing to control for the influence of individual participant traits. This efficient design allowed us to identify differences in trainee performance across conditions and highlighted a need for research that specifically controls for trainee expertise as well as their cognitive and self-regulatory abilities. Furthermore, the controlled nature of this study helped to identify potentially effective and ineffective approaches to feedback design, laying the foundation for future longitudinal research that could provide insight into the retention and transfer of acquired skills. Ideally, such work would include tasks that allow for the assessment of both near and far transfer.

8 Conclusion

We conducted a within-subjects study examining different feedback conditions to better understand how concurrent feedback can be used to improve pilot-trainee performance. We found that highly elaborative feedback that modelled experts supported training in this complex, simulator-based environment that affords the provisioning of concurrent feedback to trainees. We leveraged this ability to shed light on the interaction between trainee performance and different characteristics of concurrent feedback that include whether it is ever present or triggered by learner performance. We found that ever-present concurrent feedback placed less extraneous cognitive load on trainees, potentially allowing them to focus on learning. Specifically, the feedback that continuously modelled expert behaviour and allowed the explicit comparison of trainee behaviour to that of an expert model best supported performance. Participant reports and variability in both their performance and preferences across feedback conditions highlighted the need to adapt feedback based on individual characteristics, such as prior knowledge. This need to model expert performance and adapt feedback approaches within simulation-based learning environments expands our knowledge of how to provide concurrent feedback.

Acknowledgments

This work was conducted collaboratively among researchers from Illumia Labs and researchers in the Intelligent Robot Learning (IRL) Lab and EdTeKLA Research Group at the University of Alberta. The work was supported in part by grants from a Canada CIFAR AI Chair, Amii; Mitacs; and the Natural Sciences and Engineering Research Council of Canada (NSERC).

References

[1] 2021. *Airplane Flying Handbook (FAA-H-8083-3C)*. Technical Report FAA-H-8083-3C. United States Department of Transportation, Federal Aviation Administration, Airman Testing Standards Branch, Oklahoma City, OK. 406 pages. https://www.faa.gov/regulations_policies/handbooks_manuals/aviation/airplane_handbook

[2] J Andrew Bagnell and Jeff G Schneider. 2001. Autonomous helicopter control using reinforcement learning policy search methods. In *Proceedings 2001 ICRA. IEEE International Conference on Robotics and Automation (Cat. No. 01CH37164)*.

[3] Alok Baikadi, Carrie Demmans Epp, and Christian D. Schunn. 2018. Participating by activity or by week in MOOCs. *Information and Learning Science* 119, 9/10 (2018), 572–585. https://doi.org/10.1108/ILS-04-2018-0033

[4] Albert Bandura. 1986. *Social foundations of thought and action: a social cognitive theory.* Prentice-Hall, Englewood Cliffs, NJ.

[5] D. Barrow, Antonija Mitrovic, Stellan Ohlsson, and M. Grimley. 2008. Assessing the Impact of Positive Feedback in Constraint-Based Tutors. In *Intelligent Tutoring Systems (ITS)*. Springer, Montreal, Canada, 250–259.

[6] Joseph E. Beck and Yue Gong. 2013. Wheel-Spinning: Students Who Fail to Master a Skill. In *Artificial Intelligence in Education (Lecture Notes in Computer Science)*, H. Chad Lane, Kalina Yacef, Jack Mostow, and Philip Pavlik (Eds.). Springer, Berlin, Heidelberg, 431–440. https://doi.org/10.1007/978-3-642-39112-5_44

[7] K. Blair, D. Schwartz, G. Biswas, and K. Leelawong. 2006. Pedagogical Agents for Learning by Teaching: Teachable Agents. *Educational Technology & Society, Special Issue on Pedagogical Agents* (2006), 56–61.

[8] Robert Bodily, Judy Kay, Vincent Aleven, Ioana Jivet, Dan Davis, Franceska Xhakaj, and Katrien Verbert. 2018. Open learner models and learning analytics dashboards: a systematic review. In *Proceedings of the 8th International Conference on Learning Analytics and Knowledge (LAK)*. ACM Press, 41–50. https://doi.org/10.1145/3170358.3170409

[9] Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. *Qualitative Research in Psychology* 3, 2 (Jan. 2006), 77–101. https://doi.org/10.1191/1478088706qp063oa

[10] Susan Bull and Judy Kay. 2016. SMILI: a Framework for Interfaces to Learning Data in Open Learner Models, Learning Analytics and Related Fields. *International Journal of Artificial Intelligence in Education* 26, 1 (March 2016), 293–331. https://doi.org/10.1007/s40593-015-0090-8

[11] Deborah L. Butler and Philip H. Winne. 1995. Feedback and Self-Regulated Learning: A Theoretical Synthesis. *Review of Educational Research* 65, 3 (Sept. 1995), 245–281. https://doi.org/10.3102/00346543065003245

[12] Minghao Cai, Gokce Akcayir, and Carrie Demmans Epp. 2021. Exploring Augmented Reality Games in Accessible Learning: A Systematic Review. arXiv, Yokohomo, Japan [online], 11. http://arxiv.org/abs/2111.08214 arXiv: 2111.08214.

[13] Minghao Cai, Carrie Demmans Epp, and Tahereh Firoozi. 2022. Complex Learning Environments: Tensions in Student Perspectives that Indicate Competing Values. In *Artificial Intelligence in Education. Posters and Late Breaking Results, Workshops and Tutorials, Industry and Innovation Tracks, Practitioners' and Doctoral Consortium*, Maria Mercedes Rodrigo, Noburu Matsuda, Alexandra I. Cristea, and Vania Dimitrova (Eds.). Vol. 13356. Springer International Publishing, Cham, 144–149. https://doi.org/10.1007/978-3-031-11647-6_25 Series Title: Lecture Notes in Computer Science.

[14] Minghao Cai, Genaro Rebolledo Mendez, Gisele Arevalo, Sin Sze Tang, Yalmaz Ali Abdullah, and Carrie Demmans Epp. 2024. Toward Supporting Adaptation: Exploring Affect's Role in Cognitive Load when Using a Literacy Game. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. ACM, Honolulu HI USA, 1–17. https://doi.org/10.1145/3613904.3642150

[15] Minghao Cai, Bin Zheng, and Carrie Demmans Epp. 2022. Towards Supporting Adaptive Training of Injection Procedures: Detecting Differences in the Visual Attention of Nursing Students and Experts. In *Proceedings of the 30th ACM Conference on User Modeling, Adaptation and Personalization*. ACM, Barcelona Spain, 286–294. https://doi.org/10.1145/3503252.3531302

[16] Transport Canada. 2017. Flight Instructor Guide — Aeroplane — TP 975. https://tc.canada.ca/en/aviation/publications/flight-instructor-guide-aeroplane-tp-975

[17] Transport Canada. 2023. Flight Test Guide - Private Pilot Licence - Aeroplane - TP 13723E. https://tc.canada.ca/en/aviation/publications/flight-test-guide-private-pilot-licence-aeroplane-tp-13723e

[18] Olga Chernikova, Nicole Heitzmann, Matthias Stadler, Doris Holzberger, Tina Seidel, and Frank Fischer. 2020. Simulation-Based Learning in Higher Education: A Meta-Analysis. *Review of Educational Research* 90, 4 (Aug. 2020), 499–541. https://doi.org/10.3102/0034654320933544 Publisher: American Educational Research Association.

[19] Jeong-Im Choi and Michael Hannafin. 1995. Situated Cognition and Learning Environments: Roles, Structures, and Implications for Design. *Educational Technology Research and Development* 43, 2 (1995), 53–69. http://www.jstor.org/stable/30220993

[20] John W. Creswell and Vicki L. Plano Clark. 2011. *Designing and conducting mixed methods research* (2nd ed ed.). SAGE Publications, Los Angeles. OCLC: ocn558676948.

[21] Walter D. Davis. 2005. The Interactive Effects of Goal Orientation and Feedback Specificity on Task Performance. *Human Performance* 18, 4 (Oct. 2005), 409–426. https://doi.org/10.1207/s15327043hup1804_7 Publisher: Routledge.

[22] Stefan De Groot, Joost C. F. De Winter, José Manuel López García, Max Mulder, and Peter A. Wieringa. 2011. The Effect of Concurrent Bandwidth Feedback on Learning the Lane-Keeping Task in a Driving Simulator. *Human Factors: The Journal of the Human Factors and Ergonomics Society* 53, 1 (Feb. 2011), 50–62. https://doi.org/10.1177/0018720810393241

[23] Carrie Demmans Epp and Susan Bull. 2015. Uncertainty representation in visualizations of learning analytics for learners: Current approaches and opportunities. *IEEE Transactions on Learning Technologies* 8, 3 (2015), 242–260. https://doi.org/10.1109/TLT.2015.2411604

[24] Carrie Demmans Epp, Susan Bull, and Matthew D. Johnson. 2014. Visualising uncertainty for open learner model users. In *CEUR Proceedings of User Modeling, Adaptation and Personalization (UMAP)*, Vol. 1181. Aalborg, Denmark, 9–12. http://ceur-ws.org/Vol-1181/umap2014_poster_03.pdf

[25] Carrie Demmans Epp, Joe Horne, Britney B. Scolieri, Irene Kane, and Amy S. Bowser. 2018. PsychOut! a Mobile App to Support Mental Status Assessment Training. In *European Conference on Technology Enhanced Learning (EC-TEL): Lifelong Technology-Enhanced Learning*, Viktoria Pammer-Schindler, Mar Pérez-Sanagustín, Henrik Drachsler, Raymond Elferink, and Maren Scheffel (Eds.), Vol. 11082. Springer International Publishing, Cham, 216–230. https://doi.org/10.1007/978-3-319-98572-5_17

[26] Carrie Demmans Epp and Gordon I. McCalla. 2011. ProTutor: Historic open learner models for pronunciation tutoring. In *Artificial Intelligence in Education (AIED) (Springer Lecture Notes in Computer Science, Vol. 6738)*, G. Biswas, S. Bull, J. Kay, and A. Mitrovic (Eds.). Springer, Auckland, New Zealand, 441–443. https://doi.org/10.1007/978-3-642-21869-9_63

[27] Carrie Demmans Epp, Krystle Phirangee, Jim Hewitt, and Charles A. Perfetti. 2020. Learning management system and course influences on student actions and learning experiences. *Educational Technology, Research and Development (ETRD)* 68, 6 (Dec. 2020), 3263–3297. https://doi.org/10.1007/s11423-020-09821-1

[28] Florian Diller, Gerik Scheuermann, and Alexander Wiebel. 2024. Visual Cue Based Corrective Feedback for Motor Skill Training in Mixed Reality: A Survey. *IEEE Transactions on Visualization and Computer Graphics* 30, 7 (July 2024), 3121–3134. https://doi.org/10.1109/TVCG.2022.3227999 Conference Name: IEEE Transactions on Visualization and Computer Graphics.

[29] Holmes Finch. 2022. An Introduction to the Analysis of Ranked Response Data. *Practical Assessment, Research & Evaluation* 27 (April 2022). https://eric.ed.gov/?id=EJ1343346 Publisher: Center for Educational Assessment ERIC Number: EJ1343346.

[30] Jon Fraenkel and Bernard Grofman. 2014. The Borda Count and its real-world alternatives: Comparing scoring rules in Nauru and Slovenia. *Australian Journal of Political Science* 49, 2 (April 2014), 186–205. https://doi.org/10.1080/10361146.2014.900530 Publisher: Routledge.

[31] Bo Fu, Angelo Ryan Soriano, Kayla Chu, Peter Gatsby, and Nicolas Guardado. 2024. Modelling Visual Attention for Future Intelligent Flight Deck - A Case Study of Pilot Eye Tracking in Simulated Flight Takeoff. In *Adjunct Proceedings of the 32nd ACM Conference on User Modeling, Adaptation and Personalization*. ACM, Cagliari Italy, 170–175. https://doi.org/10.1145/3631700.3664871

[32] Kallirroi Georgila, Mark G Core, Benjamin D Nye, Shamya Karumbaiah, Daniel Auerbach, and Maya Ram. 2019. Using reinforcement learning to optimize the policies of an intelligent tutoring system for interpersonal skills training. In *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems*.

[33] Michael Guevarra, Srijita Das, Christabel Wayllace, Carrie Demmans Epp, Matthew Taylor, and Alan Tay. 2023. Augmenting Flight Training with AI to Efficiently Train Pilots. *Proceedings of the AAAI Conference on Artificial Intelligence* 37, 13 (2023), 16437–16439. https://doi.org/10.1609/aaai.v37i13.27071 Number: 13.

[34] Gerald S. Hanna. 1976. Effects of Total and Partial Feedback in Multiple-Choice Testing Upon Learning. *The Journal of Educational Research* 69, 5 (Jan. 1976), 202–205. https://doi.org/10.1080/00220671.1976.10884873 Publisher: Routledge _eprint: https://doi.org/10.1080/00220671.1976.10884873.

[35] Rory Heilakka. [n. d.]. The Airline Pilot Shortage Will Get Worse. https://www.oliverwyman.com/our-expertise/insights/2022/jul/airline-pilot-shortage-will-get-worse.html

[36] Ashley Hill, Antonin Raffin, Maximilian Ernestus, Adam Gleave, Anssi Kanervisto, and Quentin Gallouédec. 2021. Imitation Learning — Stable Baselines3 2.6.0 documentation. https://stable-baselines3.readthedocs.io/en/master/guide/imitation.html#imitation-learning

[37] Sture Holm. 1979. A Simple Sequentially Rejective Multiple Test Procedure. *Scandinavian Journal of Statistics* 6, 2 (1979), 65–70. https://www.jstor.org/stable/4615733 Publisher: [Board of the Foundation of the Scandinavian Journal of Statistics, Wiley].

[38] D.M. Hoska. 1993. Motivating learners through CBI feedback: Developing a positive learner perspective. In *Interactive instruction and feedback*, J.V. Dempsey and G.C. Sales (Eds.). Educational Technology Publications, Englewood Cliffs, N.J., 105–132.

[39] Michaël Huet, David M. Jacobs, Cyril Camachon, Cedric Goulon, and Gilles Montagne. 2009. Self-Controlled Concurrent Feedback Facilitates the Learning of the Final Approach Phase in a Fixed-Base Flight Simulator. *Human Factors* 51, 6 (Dec. 2009), 858–871. https://doi.org/10.1177/0018720809357343 Publisher: SAGE Publications Inc.

[40] Alboukadel Kassambara. 2023. rstatix: Pipe-Friendly Framework for Basic Statistical Tests. https://cran.r-project.org/web/packages/rstatix/index.html

[41] Roy P. C. Kessels, Martine J. E. van Zandvoort, Albert Postma, L. Jaap Kappelle, and Edward H. F. de Haan. 2000. The Corsi Block-Tapping Task: Standardization and Normative Data. *Applied Neuropsychology* 7, 4 (Dec. 2000), 252–258. https://doi.org/10.1207/S15324826AN0704_8 Publisher: Routledge _eprint: https://doi.org/10.1207/S15324826AN0704_8.

[42] Kenneth R. Koedinger, Albert T. Corbett, and Charles Perfetti. 2012. The knowledge-learning-instruction framework: bridging the science-practice chasm to enhance robust student learning. *Cognitive Science* 36, 5 (July 2012), 757–798. https://doi.org/10.1111/j.1551-6709.2012.01245.x

[43] James A. Kulik and Chen-Lin C. Kulik. 1988. Timing of Feedback and Verbal Learning. *Review of Educational Research* 58, 1 (March 1988), 79–97. https://doi.org/10.3102/00346543058001079 Publisher: American Educational Research Association.

[44] Thibaut Le Naour, Ludovic Hamon, and Jean-Pierre Bresciani. 2019. Superimposing 3D Virtual Self + Expert Modeling for Motor Learning: Application to the Throw in American Football. *Frontiers in ICT* 6 (Aug. 2019). https://doi.org/10.3389/fict.2019.00016 Publisher: Frontiers.

[45] Nancy L. Leech and Anthony J. Onwuegbuzie. 2007. A Typology of Mixed Methods Research Designs. *Quality & Quantity* 43, 2 (March 2007), 265–275. https://doi.org/10.1007/s11135-007-9105-3

[46] Levi H.S. Lelis, João G.G.V. Nova, Eugene Chen, Nathan R. Sturtevant, Carrie Demmans Epp, and Michael Bowling. 2022. Learning Curricula for Humans: An Empirical Study with Puzzles from The Witness. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence*. International Joint Conferences on Artificial Intelligence Organization, Vienna, Austria, 3877–3883. https://doi.org/10.24963/ijcai.2022/538

[47] Jimmie Leppink, Fred Paas, Cees P. M. Van der Vleuten, Tamara Van Gog, and Jeroen J. G. Van Merriënboer. 2013. Development of an instrument for measuring different types of cognitive load. *Behavior Research Methods* 45, 4 (Dec. 2013), 1058–1072. https://doi.org/10.3758/s13428-013-0334-1

[48] Amna Liaqat, Cosmin Munteanu, and Carrie Demmans Epp. 2020. Collaborating with Mature English Language Learners to Combine Peer and Automated Feedback: a User-Centered Approach to Designing Writing Support. *International Journal of Artificial Intelligence in Education* (July 2020). https://doi.org/10.1007/s40593-020-00204-4

[49] Anna Liu, Melissa Duffy, Sandy Tse, Marc Zucker, Hugh McMillan, Patrick Weldon, Julie Quet, and Michelle Long. 2023. Concurrent *versus* terminal feedback: The effect of feedback delivery on lumbar puncture skills in simulation training. *Medical Teacher* 45, 8 (Aug. 2023), 906–912. https://doi.org/10.1080/0142159X.2023.2189540

[50] Yanjin Long and Vincent Aleven. 2013. Supporting Students' Self-Regulated Learning with an Open Learner Model in a Linear Equation Tutor. In *Artificial Intelligence in Education (AIED)*, H. Chad Lane, Kalina Yacef, Jack Mostow, and Philip Pavlik (Eds.). Number 7926 in Lecture Notes in Computer Science. Springer Berlin Heidelberg, 219–228. http://link.springer.com/chapter/10.1007/978-3-642-39112-5_23

[51] Yanjin Long and Vincent Aleven. 2016. Enhancing learning outcomes through self-regulated learning support with an Open Learner Model. *User Modeling and User-Adapted Interaction* (Dec. 2016), 1–34. https://doi.org/10.1007/s11257-016-9186-6

[52] Danielle Lottridge, Mark Chignell, and Aleksandra Jovicic. 2011. Affective Interaction: Understanding, Evaluating, and Designing for Human Emotion. *Reviews of Human Factors and Ergonomics* 7, 1 (Sept. 2011), 197–217. https://doi.org/10.1177/1557234X11410385 Publisher: SAGE Publications.

[53] Kimberly N Martin and Ivon Arroyo. 2004. AgentX: Using reinforcement learning to improve the effectiveness of intelligent tutoring systems. In *International Conference on Intelligent Tutoring Systems*.

[54] Santosh A. Mathan and Kenneth R. Koedinger. 2002. An Empirical Assessment of Comprehension Fostering Features in an Intelligent Tutoring System. In *Intelligent Tutoring Systems*, Gerhard Goos, Juris Hartmanis, Jan Van Leeuwen, Stefano A. Cerri, Guy Gouardères, and Fàbio Paraguaçu (Eds.). Vol. 2363. Springer Berlin Heidelberg, Berlin, Heidelberg, 330–343. https://doi.org/10.1007/3-540-47987-2_37 Series Title: Lecture Notes in Computer Science.

[55] Eduardo F Morales and Claude Sammut. 2004. Learning to fly by combining reinforcement learning with behavioural cloning. In *Proceedings of the twenty-first international conference on Machine learning*. 76.

[56] Geoff Murray, Rory Heilakka, Daniel Rye, Jeff Green, and Lindsay Grant. 2022. *NEXT-GEN PILOTS A younger, tech-savvy generation on deck.* Technical Report. OlyverWyman. 16 pages. https://www.oliverwyman.com/our-expertise/insights/2022/nov/next-gen-pilots.html

[57] Donald A. Norman. 2005. *Emotional design: why we love (or hate) everyday things*. Basic Books, New York. OCLC: 254793649.

[58] Jason Osborne. 2002. Notes on the use of data transformations. *Practical Assessment, Research, and Evaluation* 8, 1 (Jan. 2002). https://doi.org/10.7275/4vng-5608 Number: 1 Publisher: University of Massachusetts Amherst Libraries.

[59] Antonin Raffin, Ashley Hill, Adam Gleave, Anssi Kanervisto, Maximilian Ernestus, and Noah Dormann. 2021. Stable-Baselines3: Reliable Reinforcement Learning Implementations. *Journal of Machine Learning Research* 22, 268 (2021), 1–8. http://jmlr.org/papers/v22/20-1364.html

[60] Siddharth Reddy, Anca D Dragan, and Sergey Levine. 2018. Shared autonomy via deep reinforcement learning. *arXiv preprint arXiv:1802.01744* (2018).

[61] BH Sreenivasa Sarma and Balaraman Ravindran. 2007. Intelligent tutoring systems using reinforcement learning to teach autistic students. In *International Conference on Home-Oriented Informatics and Telematics*.

[62] Valentin Schwind. [n. d.]. Balanced Latin Square Generator. https://hci-studies.org/balanced-latin-square/

[63] Valerie J. Shute. 2008. Focus on Formative Feedback. *Review of Educational Research* 78, 1 (March 2008), 153–189. https://doi.org/10.3102/0034654307313795

[64] D. Sleeman, A. E. Kelly, R. Martinak, R. D. Ward, and J. L. Moore. 1989. Studies of diagnosis and remediation with high school algebra students. *Cognitive Science* 13, 4 (Oct. 1989), 551–568. https://doi.org/10.1016/0364-0213(89)90023-2

[65] John Sweller, Paul L Ayres, and Slava Kalyuga. 2011. *Cognitive Load Theory*. Springer, New York.

[66] E. R. Thompson. 2007. Development and Validation of an Internationally Reliable Short-Form of the Positive and Negative Affect Schedule (PANAS). *Journal of Cross-Cultural Psychology* 38, 2 (March 2007), 227–242. https://doi.org/10.1177/0022022106297301

[67] Scott F. Turner, Laura B. Cardinal, and Richard M. Burton. 2017. Research Design for Mixed Methods: A Triangulation-based Framework and Roadmap. *Organizational Research Methods* 20, 2 (April 2017), 243–267. https://doi.org/10.1177/1094428115610808

[68] Fabienne M. Van der Kleij, Remco C. W. Feskens, and Theo J. H. M. Eggen. 2015. Effects of Feedback in a Computer-Based Learning Environment on Students' Learning Outcomes: A Meta-Analysis. *Review of Educational Research* 85, 4 (Dec. 2015), 475–511. https://doi.org/10.3102/0034654314564881 Publisher: American Educational Research Association.

[69] André Vandierendonck, Eva Kemps, Maria Chiara Fastame, and Arnaud Szmalec. 2004. Working memory components of the Corsi blocks task. *British Journal of Psychology* 95, 1 (2004), 57–79. https://doi.org/10.1348/000712604322779460 _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1348/000712604322779460.

[70] Kurt VanLehn. 2011. The Relative Effectiveness of Human Tutoring, Intelligent Tutoring Systems, and Other Tutoring Systems. *Educational Psychologist* 46, 4 (2011), 197–221.

[71] Catharine M. Walsh, Simon C. Ling, Charlie S. Wang, and Heather Carnahan. 2009. Concurrent Versus Terminal Feedback: It May Be Better to Wait. *Academic Medicine* 84, 10 (Oct. 2009), S54. https://doi.org/10.1097/ACM.0b013e3181b38daf

[72] Fangju Wang. 2018. Reinforcement learning in a pomdp based intelligent tutoring system for optimizing teaching strategies. *International Journal of Information and Education Technology* (2018).

[73] Sue Ellen Williams. 1997. *Teachers' Written Comments and Students' Responses: A Socially Constructed Interaction*. Technical Report. https://eric.ed.gov/?id=ED408589 ERIC Number: ED408589.

[74] Barry J. Zimmerman. 1989. Models of Self-Regulated Learning and Academic Achievement. In *Self-Regulated Learning and Academic Achievement*, Barry J. Zimmerman and Dale H. Schunk (Eds.). Springer New York, 1–25. https://doi.org/10.1007/978-1-4612-3618-4_1