frontiers | Frontiers in Virtual Reality

TYPE Original Research
PUBLISHED 08 February 2024
DOI 10.3389/frvir.2024.1322656

OPEN ACCESS

EDITED BY
Katharina Jahn,
Chemnitz University of Technology, Germany

REVIEWED BY
David Harris,
University of Exeter, United Kingdom
Diego Vilela Monteiro,
ESIEA University, France

*CORRESPONDENCE
Allison P. Anderson,
apanders@colorado.edu

†These authors share first authorship

RECEIVED 16 October 2023
ACCEPTED 19 January 2024
PUBLISHED 08 February 2024

CITATION
Verniani A, Galvin E, Tredinnick S, Putman E, Vance EA, Clark TK and Anderson AP (2024), Features of adaptive training algorithms for improved complex skill acquisition. *Front. Virtual Real.* 5:1322656. doi: 10.3389/frvir.2024.1322656

Features of adaptive training algorithms for improved complex skill acquisition

Alessandro Verniani[1†], Ellery Galvin[2†], Sandra Tredinnick[2], Esther Putman[1], Eric A. Vance[2], Torin K. Clark[1] and Allison P. Anderson[1*]

[1]Bioastronautics Laboratory, Ann & H.J. Smead Department of Aerospace Engineering Sciences, University of Colorado, Boulder, CO, United States, [2]Laboratory for Interdisciplinary Statistical Analysis, Department of Applied Mathematics, University of Colorado, Boulder, CO, United States

Training complex skills is typically accomplished by means of a trainer or mediator who tailors instruction to the individual trainee. However, facilitated training is costly and labor intensive, and the use of a mediator is infeasible in remote or extreme environments. Imparting complex skills in applications like long-duration human spaceflight, military field operations, or remote medicine may require automated training algorithms. Virtual reality (VR) is an effective, easily programmable, immersive training medium that has been used widely across fields. However, there remain open questions in the search for the most effective algorithms for guiding automated training progression. This study investigates the effects of responsiveness, personalization, and subtask independence on the efficacy of automated training algorithms in VR for training complex, operationally relevant tasks. Thirty-two subjects (16M/16F, 18–54 years) were trained to pilot and land a spacecraft on Mars within a VR simulation using four different automated training algorithms. Performance was assessed in a physical cockpit mock-up. We found that personalization results in faster skill acquisition on average when compared with a standardized progression built for a median subject (p = 0.0050). The standardized progression may be preferable when consistent results are desired across all subjects. Independence of the difficulty adjustments between subtasks may lead to increased skill acquisition, while lockstep in the progression of each subtask increases self-reported flow experience (p = 0.01), fluency (p = 0.02), and absorption (p = 0.01) on the Flow Short Scale. Data visualization suggests that highly responsive algorithms may lead to faster learning progressions and higher skill acquisition for some subjects. Improving transfer of skills from training to testing may require either high responsiveness or a standardized training progression. Optimizing the design of automated, individually adaptive algorithms around the training needs of a group may be useful to increase skill acquisition for complex operational tasks.

KEYWORDS

automated training, adaptive training, personalization, responsiveness, task integration, virtual reality, human spaceflight, remote training

1 Introduction

Training for complex tasks is critical for ensuring high performance in challenging operational environments. Automated training is required when facilitation is infeasible, such as when bringing large equipment or physical simulators is impractical (Braun and Manning, 2006; Simon et al., 2015), when a trainer or mediator cannot be spared among crew (Saluja et al., 2008; Landon et al., 2017; Robertson et al., 2020), or when communication delays render remote facilitation difficult (Love and Reagan, 2013; Kintz and Palinkas, 2016; Diamond, 2022). Such limitations are frequently encountered in high-consequence domains with large constraints on the ability to practice tasks in advance, including spaceflight, remote medicine, and military field operations.

Adaptive training algorithms alter subtask difficulty as a function of subject performance across a range of related subtasks, a process also known as dynamic difficulty adjustment (Hunicke, 2005; Zohaib, 2018; Moon and Seo, 2020). The core facet of adaptivity is the use of performance on previous trials as an input used to determine the difficulty of the successive trial by means of an algorithm. By contrast, non-adaptive algorithms hold difficulty across subtasks fixed at constant levels regardless of performance. The challenge-point framework suggests that training efficiency is increased by modulating difficulty to account for the skill level of the performer to provide optimal challenge (Guadagnoli and Lee, 2004). Automated training where task difficulty is adaptively matched to skill is known to increase engagement (Missura, 2015; Xue et al., 2017), improve training outcomes (Hunicke, 2005; Lang et al., 2018; Iván Aguilar Reyes et al., 2022), and enhance overall experience and reported stimulation (Schmidt, 1975). However, the effect of personalized rather than standardized progression on training outcomes and flow experience, which depend on how well a training system provides optimal challenge, has not been investigated (Vaughan et al., 2016; Jeelani et al., 2017; Yovanoff et al., 2017). The effect of progression personalization is important for applications that require complex training on multiple subtasks and where individual training variability may be high.

Further, although different performance thresholds for difficulty progression have been used for adaptive training (Koenig et al., 2011; Dhiman et al., 2016; Yang et al., 2016; Gabay et al., 2017), the effect of high vs. low thresholds (responsiveness) on training outcomes has not been systematically studied. Characterizing the optimal degree of responsiveness in automated training is important for reducing inefficiency, since an algorithm that is not adequately responsive requires excess trials to adjust difficulty and risks negatively affecting subject motivation. In addition, past investigations of training involving multiple subtasks have typically required that subjects train to proficiency on one task before progressing to a new one (Gagne, 1962; Rickel and Johnson, 1999 and 2010). The effect of training multiple subtasks in parallel and of independent subtask progression has not been investigated and is important for developing automated training systems which can most effectively teach subjects to perform complex, multi-part tasks.

The challenge-point framework suggests that optimal challenge occurs at the difficulty level where learning is maximized without large decrements in performance. When a training algorithm maintains this level of optimal challenge for subjects, they become engrossed in an activity and can enter a state called flow in which self-awareness is reduced and time is perceived to pass rapidly, conditions that accompany favorable learning (Csikszentmihalyi and LeFevre, 1989; Agarwal and Karahanna, 2000). The experience of flow has been shown to increase motivation and exploratory behavior and to improve training and learning outcomes in computer-mediated environments (Trevino and Webster, 1992; Ghani and Deshpande, 1993; Webster et al., 1993; Choi et al., 2007). However, it is not known how automated training systems can best detect and maintain a state of flow (Liu et al., 2005; Hamari et al., 2016; Oliveira Dos Santos et al., 2018) or how responsiveness, personalization, and subtask independence affect flow experience. Investigating the role of these features in improving subjective experiences during training is important for developing training algorithms which maximize flow and optimize learning.

The use of VR as a medium for adaptive training is an area of increasing interest. Immersive training in VR is known to be effective at imparting complex, operationally relevant skills (Thurman and Mattoon, 1994; Ng et al., 2019) and has been used for decades by entities like NASA to train astronauts (Psotka, 1995; Homan and Gott, 1996), including to perform operational tasks in neutral buoyancy (Everson et al., 2017; Sinnott et al., 2019), conduct simulated extravehicular-activity (Garcia et al., 2020), and to perform on-orbit repairs of the Hubble Space Telescope (Loftin et al., 1997). Although the transfer and equivalence to real world tasks has been demonstrated extensively (Kenyon and Afenya, 1995; Rose et al., 2000; Hamblin, 2005; Park et al., 2007; Moskaliuk et al., 2013), VR training does not confer benefits universally. For instance, higher levels of immersion may distract from learning (Jensen and Konradsen, 2017; Makransky et al., 2019) and virtual environments are less effective than real-world training for skilled sensory-motor coordination tasks (Harris et al., 2020; McAnally et al., 2022). Despite this, skill retention in both minimally and maximally immersive VR training systems (i.e., desktop vs. head-mounted display (HMD), respectively) is high for procedural skills (Farr et al., 2023), and highest for subjects who used HMDs when training to gain complex military medical skills (Siu et al., 2016). Skill acquisition is highest among those who train in VR, especially if used in concert with physical and/or haptics-mediated controls (Butt et al., 2018). Compared to large ground-based flight simulators, VR offers a lightweight, programmable, cost-effective, and easily operable alternative (Gupta et al., 2008) ideal for delivering automated training in remote environments.

We hypothesized that higher degrees of responsiveness, personalization, and subtask independence would improve skill acquisition, performance in a final assessment environment, and flow experience of subjects trained by automated algorithms. Although prior work has found adaptive training in immersive VR to be effective for simple procedural tasks (Sampayo-Vargas et al., 2013; Siu et al., 2016; Constant and Levieux, 2019), adaptive training on complex operational tasks has typically only been done in a physical environment (Gray, 2017; Plass et al., 2019). To bridge this gap, we conducted automated training in VR and selected three piloting subtasks with both motor learning and strategic decision-making components to investigate the role of algorithm features on complex, operationally relevant training.

2 Methods

In summary, the experimental data collection was completed over a period of 4 days, with 3 training sessions, each spaced 18–48 h apart (depending upon subjects' schedule availability). Each session contained 10 training trials for a total of 30 training trials across the 3 sessions. During the training sessions, subjects were trained to perform a spacecraft entry, descent, and landing (EDL) task in VR. The complex EDL task included three subtasks described in detail in section 2.2. The difficulty of each subtask was modulated depending on the algorithm to which subjects were assigned.

Subjects were randomly assigned to one of four groups, with their own algorithms that defined the automated training progression: Two-Up/One-Down with Lockstep (2U1D-L), Two-Up/One-Down Unlocked (2U1D-UL), One-Up/One-Down with Lockstep (1U1D-L), and Median Fixed Progression (MFP). Each of these algorithms is described in detail in Section 2.4. For the fourth and final session, subjects performed 10 trials of the EDL task in the Aerospace Research Simulator (AReS) physical (non-VR) cockpit mock-up (Zuzula et al., 2018), shown in Figure 2. Subjects in the AReS cockpit were assessed at a fixed difficulty level not encountered during training. This phase of the experiment was used to assess performance when executing the task after training was completed.

Subjects completed an Affect Grid before and after each session to provide information on induced changes in emotional state (Russell et al., 1989) as a result of the training and testing sessions; this survey is mentioned for completeness, but was not analyzed for this experiment. Following each session, subjects completed the Flow Short Scale (FSS) survey to provide self-reported information on wellbeing as proxies for the degree of task challenge, internalization, commitment, and motivation (Diener et al., 2010; Engeser and Rheinberg, 2008; Abuhamdeh, 2020; Roscoe and Ellis, 1990).

2.1 Subject demographics

A total of 32 subjects (16M/16F, 18–54 years, avg. 23.82 years) in good general health were recruited for participation in the study. Recruitment was primarily accomplished by means of flyers posted within the Aerospace Engineering Sciences building at the University of Colorado, Boulder. Subjects were prescreened and excluded from the study if they scored above the 90th percentile on the Motion Sickness Susceptibility Questionnaire (Reason, 1968; Golding, 1998; Golding, 2006) to avoid the potential for cyber sickness in highly susceptible individuals during VR training. Subjects were also excluded if they reported having color blindness or vision uncorrectable to 20/20 to avoid confounds surrounding variability in the perception of the primary flight displays and their indicators. Subjects were required to abstain from alcohol 6 or fewer hours prior to the study, and all subjects completed a reaction time test and a demographic survey, which included questions about prior piloting and flight experience and prior use of VR systems. These tests were designed to allow for an examination of individual variability in training and performance outcomes during statistical analysis.

TABLE 1 Allocation of subjects into four training groups with associated demographics, including sex composition, average age and age range, and number of subjects who reported previous flight experience.

Group	Sex (F)	Age	Prior flight experience
2U1D-L	4M/4	25 [22–32]	2
1U1D-L	4M/4	22 [18–35]	2
2U1D-UL	4M/4	27 [18–54]	2
MFP	4M/4	20 [18–25]	4

Subjects were randomly assigned to one of four training groups as shown in Table 1, which includes basic demographic information for each group. Random assignment to the MFP group began once testing in the 2U1D-L group was completed.

2.2 Training simulator

The VR training simulator was developed in Unity Game Engine version 2020.3.18. The EDL task required subjects to complete a sequence of activities, or *subtasks,* in each trial to land a spacecraft on the surface of Mars. The first subtask was landing site selection (LS), where subjects were presented with a topological map of the Martian surface (**1a**) and asked to pick a landing site that optimized the distance to marked points of scientific interest while avoiding hazardous (i.e., excessively steep) terrain within a time window of 8 s. Next, subjects performed the manual control (MC) subtask, where they used pitch and roll commands on a joystick to follow a guidance cue and other flight display information (Figure 1B) to navigate the spacecraft to their selected landing site with a limited amount of propellant for piloting. Once the spacecraft was piloted over the landing site, subjects transitioned to the terminal descent (TD) subtask, where they employed a hand thruster to minimize vertical descent speed on touchdown using a limited amount of propellant for landing on the surface (**1c**). Graphics from each of the three EDL subtasks are shown in Figure 1.

Each of the subtasks had 24 possible levels of difficulty (1–24). The AReS cockpit performance assessment was fixed at level 18, which corresponded to the full complex task at a challenging, but not infeasible difficulty; this level was skipped during training to ensure a novel difficulty for all subjects during the assessment. The LS subtask adjusted difficulty by varying the number of science sites and their proximity to dangerous terrain. The primary MC subtask adjustors were windspeed, which perturbed the spacecraft from the desired heading, and fuel capacity. The TD subtask varied the fuel available for the descent engine thruster.

During training trials, the simulator collected raw performance metrics. For the LS subtask, the simulator reported the sum distance between the selected landing site and the scientific sites of interest together with the sum distance for an algorithmically computed perfect landing site choice. A crash was reported if the subject failed to select a site in time, or if they selected a site on dangerous terrain. For the MC subtask, the simulator reported the root mean squared error between the guidance cues and the subject's pitch and roll commands. A crash was reported if the subject ran out of fuel;

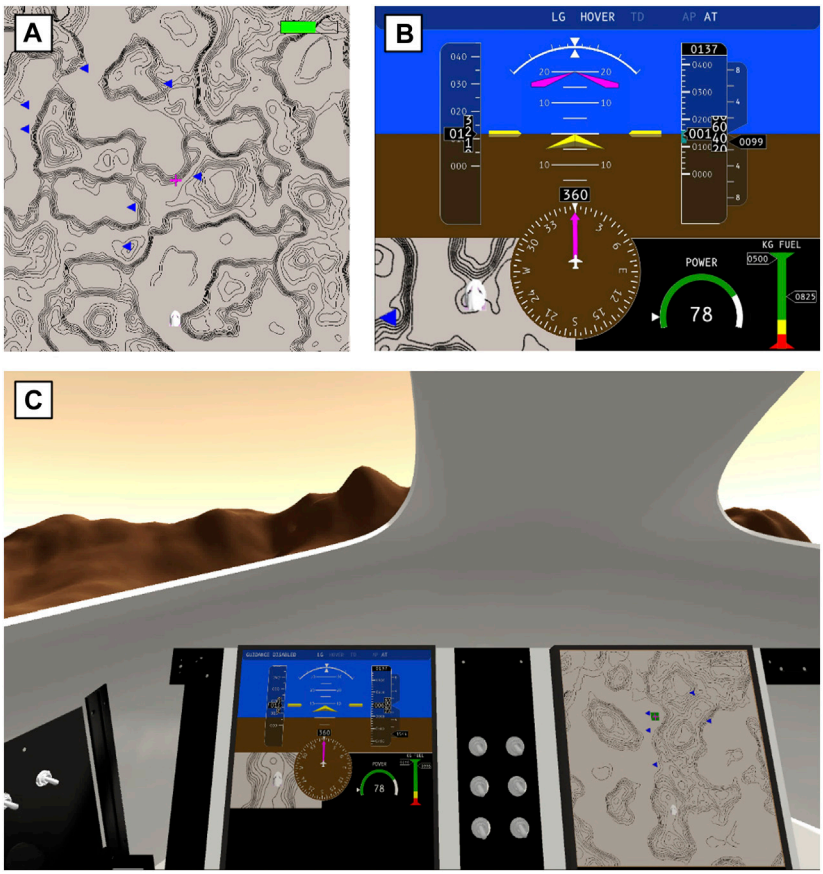

FIGURE 1
(A) Topological map on the secondary flight display used to select a landing site during the LS subtask; **(B)** Primary flight display with guidance cue, altimeter, velocity meter, fuel gauges, flight vector indicator, and mini-map for piloting during the MC subtask; **(C)** View of the virtual cockpit with primary and secondary flight displays and auto-generated Martian landscape for landing burn execution during the TD subtask.

when MC crashes occurred, the simulator was unable to score the LS and TD tasks and these data points were excluded from analyses. For the TD subtask, the simulator reported vertical speed upon contact with the ground. A crash was reported if the subject impacted the surface at a speed exceeding limits for survival.

At the end of each training trial, the simulator presented simple performance grades to the subject for each subtask: "Poor," "Adequate," or "Excellent." These grades were calculated based upon the raw performance metrics, and the thresholds for each category were determined via pilot testing prior to the experiment. For the MC subtask, the thresholds were dependent upon difficulty level because increased wind disturbance increases the root mean squared error, even under exceptional reaction times. Subjects were instructed to attempt to earn "Excellent" grades for all three subtasks.

A head-mounted display (HMD, HTC Vive Pro) was used to project the simulated interior of the AReS spacecraft cockpit mock-up to subjects during training (Sherman et al., 2023). The virtual displays and cockpit environment were designed to emulate those of AReS, the physical cockpit mock-up seen in Figure 2. Subjects used a physical joystick and hand-thruster (Logitech X-52 Saitek X52 Pro Flight System) to perform tasks, and physical inputs to both controllers were recorded in a server in addition to performance data.

2.3 Automated training algorithms

Four adaptive difficulty training algorithms were developed, visualized in Figure 3. The first adaptive algorithm takes the form of Two-Up/One-Down (2U1D), which we use as our baseline. In this algorithm, difficulty is quantized and can both ascend (Up) and descend (Down) by increments of 1 difficulty level based upon performance on preceding trials. This staircase is modeled on the parameter estimation by sequential testing (PEST) method often used for signal detection tests, whereby the strength of a signal is diminished after successive correct detections of a stimulus (Taylor and Creelman, 1967; Pollack, 1968; Leek, 2001). Requiring a higher number of correct detections increases fidelity, but with diminishing effect (Levitt, 1971). Thus, when applied to training paradigms, subjects in the 2U1D staircase were required to score "Excellent" performance grades on a subtask twice at the same level of difficulty in succession (Two-Up) before that difficulty was modulated up by one level. Conversely, subjects who scored a "Poor" performance grade just

10.3389/frvir.2024.1322656

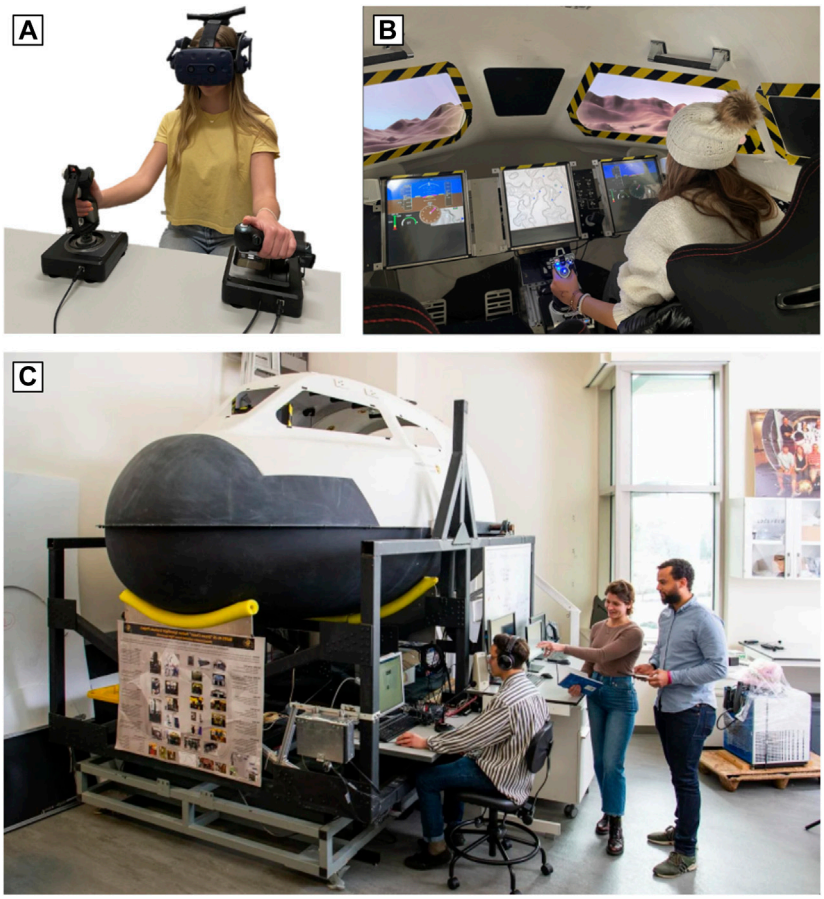

FIGURE 2
(A) Head-mounted VR display, joystick, and hand-thruster used by subjects during virtual training **(B)** AReS cockpit interior during EDL scenario testing, with physical flight displays and identical flight controls visible **(C)** View of the AReS cockpit exterior in the Bioastronautics Laboratory at the University of Colorado, Boulder, with external monitors and controls visible.

once on a subtask (One-Down) would have the algorithm modulate the difficulty down by one level for that subtask.

In contrast to the baseline 2U1D progression algorithm, we investigated the role of responsiveness by developing a second algorithm that uses a One-Up/One-Down (1U1D) variant which requires only a single, rather than double, excellent performance to increase subtask difficulty. Both forms of the staircase have been used in prior work, with 2U1D being employed for rehabilitation training in virtual environments, and its 1U1D counterpart being used *de facto* for neurorehabilitation, balance and gait training, and training of fine motor movements (Cameirão et al., 2010; Grimm et al., 2016; Gray, 2017; Kumar et al., 2018; Saurav et al., 2018). By comparing the outcomes on skill acquisition, transfer, and performance outcomes of 2U1D *versus* 1U1D, we were able to isolate the effect of heightened sensitivity and responsiveness to subject performance.

We also investigated the role of independent *versus* locked subtask progression on learning outcomes. To do so, we created a system called "lockstep" such that when difficulty levels begin to diverge, the difficulty of the highest-level subtasks were "locked", unable to increase (regardless of performance) until the difficulty of the lowest level subtask improved to within one difficulty level. This lockstep was intended to prevent the development of one skill at the expense of mastering the entire integrated task. By contrast, an unlocked staircase allowed for independent modulation of subtask difficulty. Staircase independence assumes that though sequential and interconnected, subtasks require disparate skills and that subjects may develop skills faster on certain subtasks compared to others. When the 2U1D staircase was unlocked (2U1D-UL), there were three staircases functionally operating in parallel, modulating difficulty according to subject performance for individual subtasks. This allowed for unconstrained progression on the basis that subjects will learn more effectively at varying levels of challenge across subtasks.

Finally, we investigated the role of personalization by developing the Median Fixed Progression (MFP) algorithm, a non-adaptive, fixed progression based on the median difficulty progression on each subtask by subjects in the baseline 2U1D condition with lockstep enabled (2U1D-L). MFP mimics the progression characteristics of adaptivity without responding to individual subject performance. It isolates the effect of subject-specific adaptivity on those whose performance, and thus training needs, differ from the median, either because of exceptional ability or unique difficulty in skill acquisition. The MFP condition captured the initial decline in

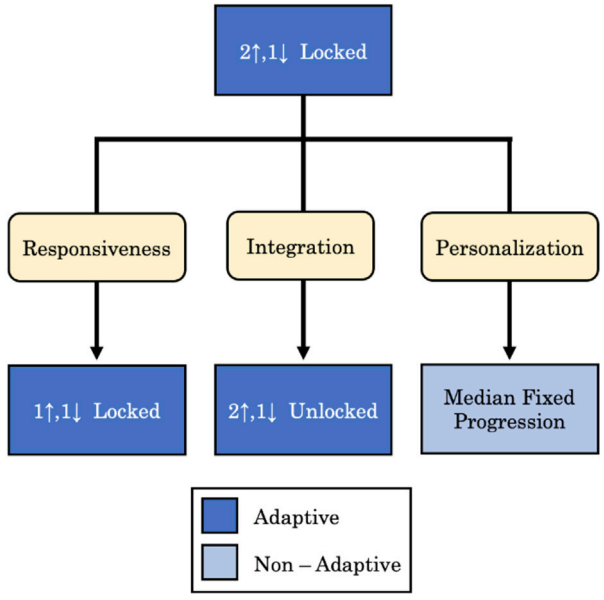

FIGURE 3
Features of automated algorithms investigated in the study: responsiveness, by changing the progression sensitivity; integration, by removing subtask lockstep; and personalization, by creating difficulty fixed to the median progression of adaptively trained subjects.

difficulty across subtasks as subjects familiarized themselves with subtasks and associated controls, as well as the eventual and gradual increase in difficulty as subjects become familiar with controls and begin honing specific motor skills.

2.4 Statistical methods

2.4.1 Data preparation

The raw performance metrics collected by the simulator were transformed into a *subtask skill metric* for each subtask and a summary mission success indicator for the trial as a whole. The subtask skill metrics—prior to difficulty control, as discussed below—are continuous values between zero and one such that zeros correspond to crashes, values less than 0.25 are "Poor" grades, values between 0.25 and 0.75 are "Adequate" grades, and values greater than "0.75" are "Excellent" grades for the subtask. This standardization ensures that the subtask skill metrics are comparable across subtasks. To construct these numbers, we computed a linear transformation of the raw performance metrics, and then applied the error function. The transformation was chosen such that the result of the error function corresponds with the correct performance grades. We refer to the *integrated skill* on a given trial to be the sum of the three subtask skill metrics. The *mission success indicator* takes the value one if the performance grades contain at least one "Excellent" grade and no "Poor" grades, and zero otherwise.

A difficulty control mechanism was used to account for the differences between achieving an "Excellent" grade at each difficulty level. Conceptually, we considered an "Excellent" grade on the highest difficulty level of 24 to correspond with having mastered the task at 100% of the possible difficulty. An "Excellent" grade on level 23 corresponds with a 23/24 proportion of the task's possible difficulty. As such, we multiplied the subtask skill metric by the difficulty level at which it was completed, divided by 24. This adjustment means that the maximum possible subtask skill metric at the starting level of 12 is 0.5, and the corresponding maximum integrated skill at level 12 is 1.5 (if all three subtask skill metrics were at their maximum of 0.5).

2.4.2 Automated training

The difficulty staircases were analyzed to assess how the algorithms facilitated difficulty progression during training. We then compared the staircases with performance measured in the AReS mock-up to assess whether the facilitation affected task execution after training. To determine whether the difficulty staircases differed between groups, we used Mixed-ANOVA (Kassambara, 2023) to compare each subject's median difficulty level attained between groups and within each subject's sessions. Since the sphericity assumption was violated (Mauchly: group*session interaction $p < 0.00005$) due to the diffusion of subjects' difficulty levels over time, Greenhouse-Geisser corrections were used. Pairwise t-testing with Holm's corrections were used in *post hoc* tests. To compare the staircases with performance in the mockup, we created a visualization mapping the difficulty attained on the last trial *versus* the skill level attained in the mockup (Figure 7).

We hypothesized that the subject's skill level progression would initially increase quickly before plateauing. To characterize this learning process in training, nonlinear mixed effects modeling was used to fit a logistic (learning) curve for skill level as a function of trial number (x) with three parameters: the asymptote (asym), the scale (scal), and the midpoint (xmid) (Dang et al., 2017; Pinheiro et al., 2017).

$$skill = \frac{asym}{1 + e^{-\frac{\left(x - x_{mid}\right)}{scal}}} \quad (1)$$

The asymptote describes the maximum achieved skill. The scale describes the timescale for attaining the asymptote, and the midpoint describes the inflection point on the curve. We used fixed effects for the asymptote and scale by group, with a global midpoint. Random effects were used for the asymptote by subject. The residuals in the final model were skewed slightly left (as ascertained from a QQ plot), and their variance was not homogeneous in time because subjects' skills diverge as their difficulty levels diverge. Consequently, measurements taken near the end of training were too highly penalized for large residuals, and some outlying negative residuals (crashes) were strongly penalized. These effects caused underestimates of the asymptote, but the result was uniform across groups, retaining cross-group comparability.

To assess whether the maximum achieved skill during training and the learning timescale differed between groups. A permutation test was used to ascertain whether the fixed effects modeling parameters differed significantly from the 2U1D-L baseline. The null hypothesis was that the training algorithm had no effect on the measurement of the asymptote fixed effects. The same model was fit for 1,000 randomized training algorithm assignments to produce a sampling distribution for the differences between the asymptote fixed effects (Finos et al., 2013). We then compared the sampling distribution with the differences obtained for the real data to produce a *p*-value.

2.4.3 Cockpit performance

We aimed to ascertain how differences in training groups corresponded with differences in performance during cockpit evaluation. Our hypothesis was that training groups attaining higher median difficulty levels, or which achieved higher asymptotic skill levels in training, would proceed to perform better during evaluation. We compared the composite and integrated skill metrics attained by subjects in each group when evaluated on the first trial in the AReS cockpit mock-up. We limited our evaluation to the first trial to preclude measuring any potential skill acquisition and adaptation to the AReS environment as a result of repeated trials. The first trial is especially relevant as it is analogous to a pilot's first transition from simulator training to physical execution. Depending on assumption validity, we used either one-way ANOVA or Kruskal–Wallis to compare composite and integrated skill metrics as assessed on the first trial in AReS. Where significant results were obtained, pairwise t-tests factor or Dunn's tests were used for *post hoc* analysis with Holm's correction.

The mission success indicator provides a summary of whether a given trial would have resulted in a successful mission. We aimed to test whether the training condition affected the probability of mission success. To compare mission success probability for each subtask across groups, we modeled successes on the first trial in the AReS mockup as draws from one of four Bernoulli distributions such that each group is permitted its own success probability. A generalized likelihood ratio test was conducted to test the null hypothesis that these success probabilities were equal (Amsalu et al., 2019).

2.4.4 Subjective perceptions

Differences in self-reported perceptions of flow were analyzed between groups during training. Parametric tests were exchanged for nonparametric tests when assumptions were violated. For a subset of attributes from the Flow Short Scale survey (i.e., flow experience, perceived task importance, fluency, and absorption), we used mixed ANOVA on the training data to compare between-subject training conditions and within-subject sessions. Where significant results were obtained, pairwise t-tests or Dunn's tests were used for *post hoc* analysis with Holm's correction.

Mixed ANOVA was used for analysis as the method has been widely employed in studies that assess skill transfer from virtual to physical environments, as well as to assess attributes of the Flow Short Scale surveys (Wilbert et al., 2010; Francis et al., 2020; Cooper et al., 2021; King et al., 2022). All analyses were performed using R version 4.3.0 (R Core Team, 2023) and Python 3.10.6 Jones et al., 2001; McKinney, 2010; Vallat, 2018; Wickham et al., 2019.

3 Results

3.1 Automated training

Difficulty progressions of each automated training group as a function of subtask and training trial are shown in Figure 4. The LS task (Figure 4-LS) was the easiest subtask. Subjects in all algorithms showed the ability to perform the task well at the entry level

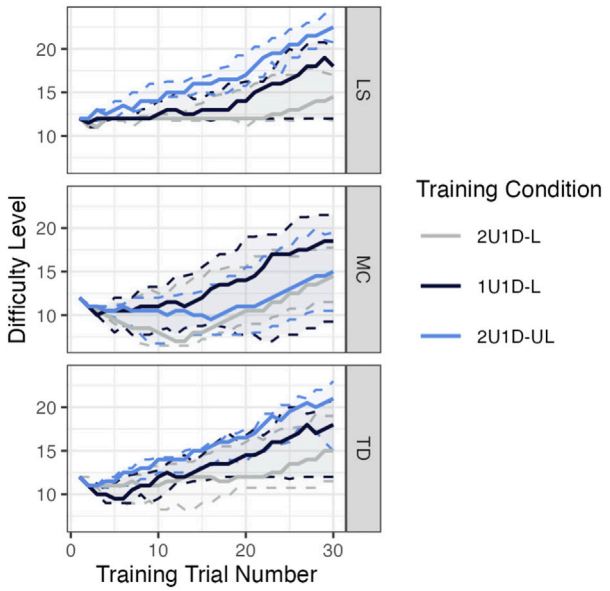

FIGURE 4
Solid lines show the median difficulty level (*y*-axis) attained by subjects in each group (color) and each task (paneling) as a function of training trial number (*x*-axis). Colored dashed lines above and below each corresponding solid line represent the 0.75 and 0.25 quantiles respectively. The MFP group is omitted because its difficulty progression is defined as the 2U1D-L median.

difficulty, level 12. Over the 30 training trials, median difficulty level achieved diverges. Testing the interaction between training session and group yielded significance ($p = 0.027$). The MC subtask (Figure 4-MC) was the most difficult of the three task dimensions, as shown by the initial decrease in difficulty as all algorithms stepped down as a result of poor performance. This dip in difficulty level at the start of training is caused by task familiarization and was sustained for many subjects during all training trials.

The 1U1D-L group tends to excel faster on the MC subtask, but the interaction between training session and group did not yield significance ($p = 0.733$), likely due to high individual variability. The TD subtask (Figure 6-TD) also showed a familiarization dip in the initial trials, but subjects then began to progress in difficulty over the remaining training trials. While the 2U1D-UL group reached the highest difficulty level, the interaction between training session and group did not reach significance ($p = 0.199$). Paired t-testing with Holm's corrections found no significant differences in the difficulties achieved by each group in any session; we did not assess contrasts between sessions of any one group because these comparisons are not of research interest. See Table 2 for complete results.

Integrated skill acquisition is shown below in Figure 5. All training conditions show subjects improving over time. Some high performing subjects improve near linearly, while other subjects follow the learning curves more closely. The 2U1D-L, 1U1D-L, and 2U1D-UL groups all have similar scale parameters, and the permutation test on these parameters found a significant difference between the MFP group and the baseline, 2U1D-L ($p = 0.005$). This difference likely occurred because the MFP reaches high difficulty levels—and thus higher integrated skill levels—slower than

TABLE 2 Mixed ANOVA results for a comparison of the median difficulty levels attained by subjects within each session and between each training condition.

Dependent variable		2U1D-L			1U1D-L			2U1D-UL						
	Training session	S1	S2	S3	S1	S2	S3	S1	S2	S3		Algorithm	Session	Interaction
LS Difficulty	Mean	0.50	0.55	0.62	0.50	0.58	0.68	0.54	0.68	0.86	F	(2.21) 5.274	(2.42) 50.784	(4.42) 3.926
	SD	0.02	0.11	0.18	0.02	0.10	0.17	0.05	0.09	0.09	pes	0.334	0.707	0.272
											p	<0.05*	<0.001***	<0.05*
MC Difficulty	Mean	0.42	0.43	0.56	0.44	0.52	0.62	0.43	0.42	0.56	F	(2.21) 0.261	(2.42) 13.703	(4.42) 0.349
	SD	0.08	0.21	0.24	0.07	0.19	0.30	0.07	0.19	0.25	pes	0.024	0.395	0.032
											p	>0.05	<0.001***	>0.05
TD Difficulty	Mean	0.45	0.47	0.56	0.45	0.56	0.67	0.50	0.59	0.74	F	(2.21) 1.571	(2.42) 35.816	(4.42) 1.792
	SD	0.07	0.21	0.23	0.06	0.12	0.16	0.04	0.11	0.17	pes	0.13	0.63	0.14
											p	>0.05	<0.001***	>0.05

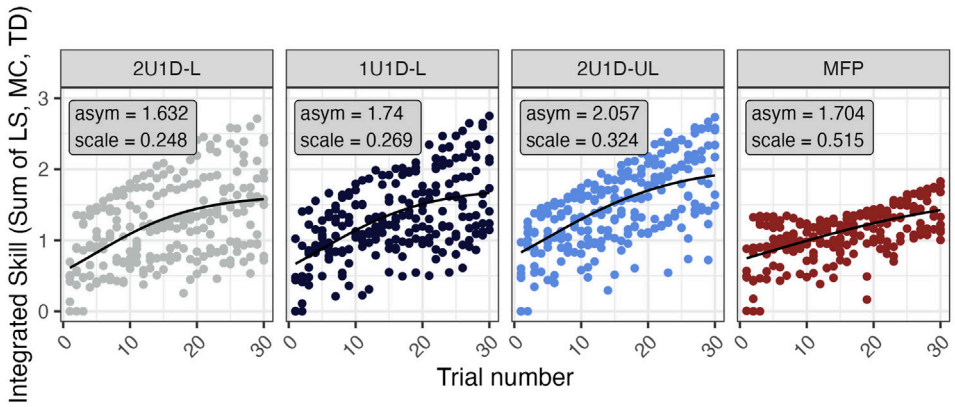

FIGURE 5
Points show each subject's integrated skill (*y*–axis) as a function of training completion (*x*-axis). The panels show each training condition in its corresponding color. The black curves show the fixed effects predictions computed via nonlinear mixed effects modeling with a learning curve. The group parameter estimates may be found inset in each panel. The global estimate for xmid is 0.160, which corresponds roughly with the sixth trial.

the highest performers in other groups. The other differences in scale parameters were not found to be significant (1U1D-L: p = 0.8211; 2U1D-UL: p = 0.4110). The 2U1D-UL group has the highest computed asymptote, likely due to its unlocked progression, but permutation testing did not find a significant difference from baseline for any group (1U1D-L: p = 0.6831, 2U1D-UL: p = 0.105, MFP: p = 0.8079).

To further investigate the effects of individual differences on the asymptote parameter, Figure 6 shows the distribution of asymptote parameters calculated for each subject. The overall standard deviation for the between-subject random effect on asymptote was 0.459, which is comparable with the mean differences between groups and may reduce power to detect group level effects. The 2U1D-L group exhibits the greatest variation in subjects' asymptotes, while the MFP group had much smaller differences. Fixed progression may result in more consistent training outcomes, while adaptive progressions seem to enable higher performance for some subjects.

3.2 Cockpit performance

To investigate the connection between the learning process and evaluation in the physical AReS mockup, we examined whether training at a high difficulty level predicted excellent performance in the cockpit mockup. The difficulty level attained in training was a strong predictor of the subtask skill metric in the physical AReS mockup for the MC task, but not the other tasks, as shown in Figure 7. On the MC task, all seven of the subjects who trained above the physical mock-up testing difficulty achieved a performance score of excellent, and those who trained below largely did not perform as well. For the LS task, the difficulty attained in training does not predict performance, as evidenced by the flat regression curve. The TD task shows that subjects trained above the difficulty level in the mockup had a mixture of high and low scores.

Figure 7 also highlights how the algorithms differ in skill transfer from the virtual to the physical environments. Data located above the regression lines suggests that subjects training at that difficulty went on to perform better than average in the cockpit. The 2U1D

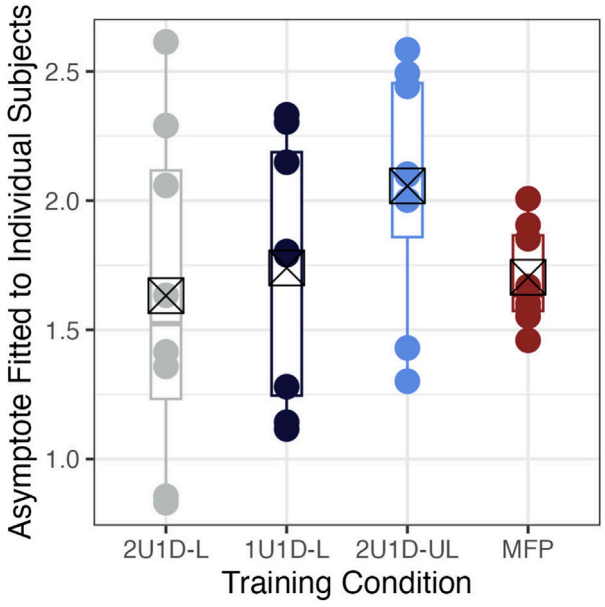

FIGURE 6
The points and summarizing box plots show the asymptote parameters (*y*-axis) fit as a random effect for each individual subject by training condition (*x*-axis). The crossed-out squares show the means for each group, which are definitionally also the corresponding fixed effect asymptote parameters.

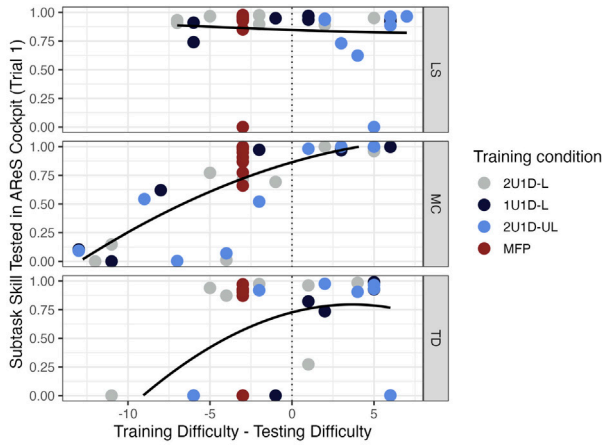

FIGURE 7
Data points describe the subtask skill metric (*y*-axis) measured on trial 1 in the physical AReS mockup as a function of the difference between the last difficulty level at which they trained and the difficulty level at which they were assessed in the mockup. The vertical gray dotted line partitions the *x*-axis into those trained below the tested difficulty (left) and those trained above. Black trend lines are Loess regressions for all data points collectively.

conditions exhibit below average transfer on all subtasks, with twice as many subjects scoring poorly relative to the difficulty at which they trained.

The 2U1D-UL group scored the lowest—or otherwise comparable—mean skill across all metrics as seen in Figure 8; compare this finding with their comparatively excellent performance in training with the NLME asymptote. Subjects trained with the 2U1D-L algorithm scored well on the LS task. The MFP group has a substantially smaller standard deviation on the MC task and notably excellent performance on the MC task as well. However, none of the hypothesis testing yielded significant results (Table 3).

When assessing whether the probability of success on the first trial in the physical AReS mockup differed between groups, we found no significant results (generalized likelihood ratio test, Bernoulli: mission success $p = 0.990$; LS failure $p = 0.773$; MC failure $p = 0.3110$; TD failure $p = 0.697$).

3.3 Subjective perception in training

Differences between automated training groups in measures of subjective experience during training are shown in Figure 9, depicting the mean flow scores for a selection of four subscales on the Flow Short Scale. We note the 2U1D-L group stands out having reported the highest flow experience and fluency in session 2 and 3, as well as the highest task importance and absorption throughout training. The four groups differed significantly in absorption (group $p = 0.013$). The interaction terms for flow experience ($p = 0.012$), fluency ($p = 0.023$), and absorption ($p = 0.013$) during training also differed. See Table 4 for full testing results. These results indicate the 2U1D-L group differed over time on these dimensions compared to the other groups.

4 Discussion

4.1 Comparison of subtasks

The MC subtask differs from the other two subtasks in several ways. Notably, Figure 4 suggests that MC is the most challenging subtask for subjects to become familiar with because the median difficulty level initially dips significantly below the starting difficulty. This is not the case for the other two subtasks. Although subjects may be scoring well on the other subtasks, training algorithms with lockstep initially prevent further difficulty progression on the LS and TD subtasks until subjects begin to score well consecutively on MC. For instance, the 2U1D-UL condition reached the highest difficulties in the LS and TD tasks because these subjects were not locked to low difficulty levels on the challenging MC task.

This higher achievement in training did not, however, correspond to better performance in testing within the physical AReS mockup (Figure 8). Since higher training difficulty did not predict higher performance in the physical AReS mockup testing (Figure 7), it is likely that either the difficulty increments were not large enough to have an effect or the skill metrics were only weakly related to how the task changes between difficulty levels. Additionally, it may be the case that subjects who train at high difficulties and leap down to the testing difficulty may face issues adapting back to the easier task. Figure 7 shows this observation seems to hold in general for the LS and TD subtasks. This suggests that skill

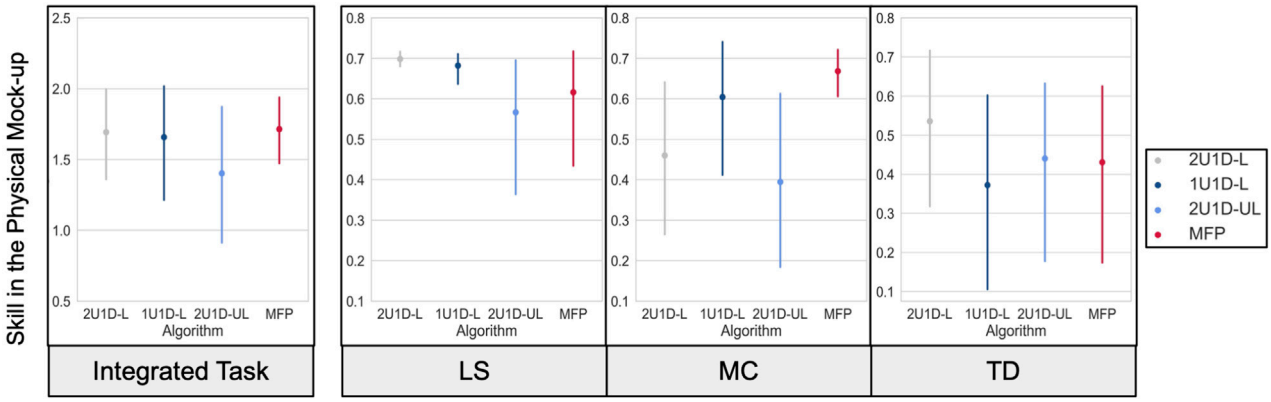

FIGURE 8
Mean skill type (*y*-axis) for each algorithm: 2U1D-L, 1U1D-L, 2U1D-UL, and MFP (*x*-axis) during the first trial in the physical mock-up, faceted by task. The error bars showcase the standard deviation for each skill type. The integrated task corresponds to the summation of LS skill, MC skill, and TD skill.

TABLE 3 ANOVA results for a comparison of the integrated skill levels attained by subjects between each training condition when evaluated on their first trial in the AReS physical mockup.

Dependent variable		2U1D-L	1U1D-L	2U1D-UL	MFP		Algorithm
Integrated skill 1st trial in AReS mockup	Mean	1.69	1.66	1.4	1.72	F	(2.21) 5.274
	SD	0.44	0.54	0.74	0.32	pes	0.334
						p	<0.05*

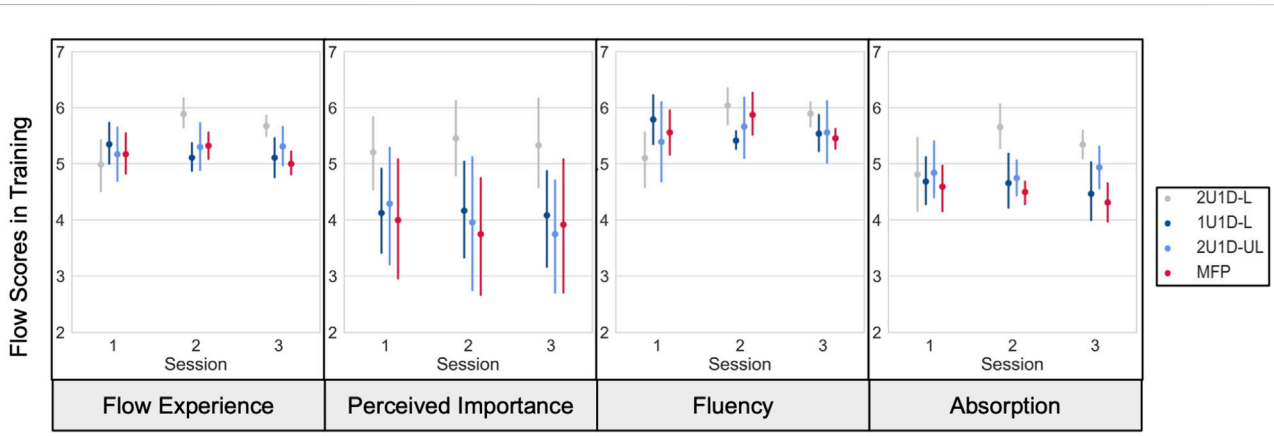

FIGURE 9
Depicts the mean (*y*-axis) at each training session (*x*-axis) for the 4 selected attributes of the Flow Short Scale survey, partitioned by each algorithm (colors). The error bars depict the standard deviation for each flow attribute.

measurements made from the MC subtask best characterize the effects of difficulty modulation, and we prioritize these results accordingly.

4.2 Responsiveness

We hypothesized that higher algorithmic responsiveness would lead to faster skill acquisition, higher tested performance, and improved self-reported measures of flow by more rapidly responding to deviations from an optimal level of challenge. Our results show that responsiveness, represented by the 1U1D-L algorithm, leads to higher learning rates, but not improved flow.

The 1U1D-L group progresses faster than the other groups on the MC subtask during training, a unique finding illustrated in Figure 4. Among the training groups, subjects in 1U1D-L also tested well on the MC subtask in the physical AReS mockup (Figure 8) because its higher adaptivity rate in training allowed access to higher

TABLE 4 ANOVA results for a comparison of Flow Short Scale responses provided by each subject within each training session and between each training condition. Only four subscales are shown.

Dependent variable		2U1D-L			1U1D-L			2U1D-UL			MFP						
		S1	S2	S3	S1	S2	S3	S1	S2	S3	S1	S2	S3		Algorithm	Session	Interaction
Flow Experience	Mean	4.99	5.89	5.68	5.35	5.11	5.11	5.18	5.3	5.31	5.18	5.32	5	F	(3.28) 1.316	(2.56) 2.32	(6.56) 3.03
	SD	0.68	0.39	0.27	0.56	0.37	0.49	0.72	0.58	0.5	0.52	0.36	0.3	η^2	0.12	0.076	0.076
														p	>0.05	>0.05	<0.05*
Perceived Importance	Mean	5.21	5.46	5.33	4.13	4.17	4.08	4.29	3.96	3.75	4	3.75	3.92	F	(3.28) 1.88	(2.56) 0.52	(6.56) 0.84
	SD	0.93	0.96	1.12	1.12	1.26	1.24	1.56	1.68	1.46	1.56	1.46	1.67	η^2	0.17	0.02	0.08
														p	>0.05	>0.05	>0.05
Fluency	Mean	5.1	6.04	5.9	5.79	5.42	5.54	5.4	5.67	5.56	5.56	5.88	5.46	F	(3.28) 0.12	(2.56) 2.42	(6.56) 2.69
	SD	0.71	0.48	0.32	0.64	0.25	0.48	1.02	0.79	0.82	0.56	0.53	0.26	η^2	0.01	0.08	0.22
														p	>0.05	>0.05	<0.05*
Absorption	Mean	4.81	5.66	5.34	5.69	4.66	4.47	4.84	4.75	4.94	4.59	4.5	4.31	F	(3.28) 4.28	(2.56) 0.67	(6.56) 1.61
	SD	0.95	0.57	0.37	0.62	0.72	0.75	0.77	0.47	0.56	0.64	0.31	0.51	η^2	0.31	0.02	0.15
														p	<0.05*	>0.05	>0.05

10.3389/frvir.2024.1322656

training difficulties relative to the other algorithms for the MC subtask specifically. We posit that the 1U1D-L algorithm's higher responsiveness best matches the effective learning rate of the subjects, while the 2U1D algorithms were too conservative in increasing difficulty.

Higher learning rates may, however, negatively impact self-reported measures of flow. The 2U1D-L baseline group generally had the highest flow scale scores including in direct contrast with the higher adaptivity rate of 1U1D-L. This observation might mean that the flow channel does not correspond with the fastest learning progression; the adaptivity rate influences flow. This may occur because highly responsive difficulty adjustments more frequently expose subjects to unduly high levels of challenge, negatively impacting performance and affecting flow by more frequently prompting negative feedback.

We propose that the appropriate degree of responsiveness may be task dependent. A difficult task with large difficulty increments may require lower responsiveness, where multiple success counts are required before increasing difficulty. On the other hand, an easier task with small increments may benefit from the accelerated progression of high responsiveness. Naturally gifted or experienced subjects may require higher algorithmic responsiveness to enable them to reach their fastest potential progression. The underlying principle is that the algorithm should adapt as quickly as the subject learns without responding to noise in the subject's performance.

4.3 Subtask integration

We hypothesized that algorithms without locked subtask integration would lead to faster skill acquisition, higher tested performance, and improved self-reported measures of flow by allowing algorithms to supply an optimal difficulty level across subtasks. Our results show that when subtask difficulty is modulated independently rather than in lockstep, progression and skill acquisition is heightened on subtasks which subjects find easier. In a highly integrated algorithm, lockstep limits the maximum skill achievement over a fixed training period since easier subtasks are held back or anchored by difficult ones.

Although integration does not seem to benefit overall performance (Figures 5, 8), it may improve self-reported measures of flow. The 2U1D-UL group does not share in the higher flow scale scores enjoyed by the 2U1D-L group (Figure 9), indicating that the lockstep affects flow. We note that in the presence of one subtask, MC, that is harder than the rest, the lockstep ensures that the subject experiences a distribution of subjective challenge levels; the hardest task prevents progress on the others, thereby enforcing that some tasks feel easier than others. This effect may play a role in the facilitation of flow.

Subtask independence may be favorable when the subtasks vary in complexity or when maximum skill achievement is desired. Lockstep may be employed when subjective perceptions are a priority. This suggests that the optimal challenge-point framework is not generalizable across subtasks if difficulty is variable between them. Applied to spaceflight, astronauts who train to perform variably complex tasks such as vehicle control, habitat maintenance, or medical operations would likely benefit from independent subtask difficulty progression.

4.4 Personalization

We hypothesized that personalized, individually adaptive algorithms, would lead to faster skill acquisition, higher tested performance, and improved self-reported measures of flow by supplying an optimal challenge-point tailored to individual needs. Our results show that personalization results in faster skill acquisition across training for some subjects, but slower skill acquisition for other subjects. On average, the timescale for skill acquisition is faster in a personalized context.

The MFP group performed strongly on the MC subtask in the physical AReS mockup testing. We note that all subjects in the MFP group were trained up to level 15, close to the testing difficulty of 18. All other groups had subjects training at a variety of difficulty levels, including some much lower than 18 (Figure 7). This suggests that the forced training up to level 15 benefited some subjects who may not have attained this difficulty level under a personalized progression. Conversely, however, the stronger subjects were held back from attaining exposure to higher difficulties.

Quantitatively, the standard deviation for skill measurements in the MFP group was generally smaller than the other groups (Figures 5, 6). This phenomenon likely occurred because all members of the MFP group trained at the same difficulty levels; the difficulty control mechanism scales skill measurements such that a wider spread in difficulty levels leads to a wider spread in skill. In addition, the permutation test on the nonlinear mixed effects model showed a significantly larger scale parameter relative to the other groups, yet we tote that the asymptote for the MFP group is in the range of that achieved in the other groups. As such, it is reasonable to speculate that the MFP group would reach the same level of proficiency if allowed to perform more training trials. We believe that the shape of this learning curve is highly dependent on the choice of fixed progression. For instance, had we chosen a single, static, low difficulty, the maximum achievable skill would be much lower—lower asymptote—and the subjects would reach this skill faster—smaller scale.

These findings imply that a standardized rather than personalized progression will result in more consistent performance at a level dependent upon the choice of progression when only a fixed number of training trials are administered. This aligns with prior findings that individually customized difficulty in training improves spatial cognitive performance (Jeun et al., 2022) and that spatial training in augmented reality is more effective when it is personalized to account for prior knowledge (Papakostas et al., 2022). This also suggests that the optimal challenge-point may not always be best acquired by tailoring a progression to individual needs during training, but rather by creating a standardized progression which reliably exposes subjects to difficulty levels akin to those of the testing environment. Our study expands upon such work by characterizing the effect of personalization on complex operational tasks and by showing that training subjects to a non-personalized adaptive progression results in tighter variance of training outcomes. The MFP forces weak subjects to encounter

difficulty levels higher than they would naturally attain benefit their overall performance. Meanwhile, gifted subjects are held back from reaching higher potential performance. Overall, the lack of personalization for an MFP results in homogenous performance at a moderate level. For applications where training time abounds, such as in isolated, confined, or extreme (ICE) environments or long-duration spaceflight, personalization is a potent method for ensuring that a subject reaches their full training potential.

4.5 Limitations

Some results could not be adequately assessed by hypothesis testing because the effect sizes were small relative to natural within and between subject variation, particularly for the skill metrics; a larger sample size may benefit power. Thus, we highlighted conclusive findings and suggest qualitative interpretation where appropriate. Furthermore, we note that the alpha level was set to 0.05 for each family of hypothesis tests. Certain limitations also arise from our cohort of participants, as a larger subject pool would have provided greater statistical power. Further, subjects in the MFP condition were 5 years younger on average than those in the baseline condition, and half of the MFP subjects reported previous flight experience, factors which may have impacted measures regarding personalization. Also, the experimental subjects were primarily comprised of aerospace engineering students. While this cohort may have been well primed to perform this simulation, as would be the case for using this training approach with an operational group of participants, these factors place potential limits on the generalizability of the results to older or non-technical populations. Although differences in the amount and frequency of positive feedback between algorithms were hypothesized to vary as a function of responsiveness, these differences were not captured explicitly and may be of interest in investigating responsiveness in future work. Finally, survey fatigue in some subjects may have eroded the significance of self-reported measures of flow. Although dividing the training across 3 days may have provided a structural mitigation to fatigue, the degree to which variability in survey motivation may have affected results could not be quantified.

5 Conclusion

Automated training algorithms with personalized, highly responsive difficulty adjustments operating independently across subtasks lead to faster learning rates and greater skill acquisition, particularly for complex, operationally relevant tasks. Our results demonstrate that personalized training is beneficial for the average subject because they are free to progress faster through training and attain the highest levels of proficiency. Meanwhile, a standardized system of adaptive progression creates more predictable and consistent training results when constrained by finitely many training trials. This finding is particularly useful for applications where a larger number of subjects of varying ability must be trained to a specific level of competence.

The optimal training algorithm depends on the goals of the training. An algorithm for fast learning among gifted subjects is one with sufficient responsiveness to match a subject's learning rate on a given task. Military settings seeking to quickly identify fast learners could train subjects with a highly responsive algorithm and assess who excels the most. When achieving flow is important, higher subtask integration may be helpful on lengthy time scales while lower subtask integration facilitates greater skill acquisition on short time scales. Secondary school settings would benefit from higher integration because motivating students and increasing task engagement is paramount. When higher consistency in performance between trainees is desired over maximizing achievement, standardization may be more appropriate than personalization. Medical settings aiming to train patients on personal medical equipment may prefer standardization because assuring average performance in a short time is often more important than enabling high performers. To facilitate transfer of skills from training to testing, it may be helpful to prioritize either higher responsiveness or less personalization. While this study was not intended to study these populations or this application, these results may serve as a starting point for future work.

Characterizing these effects is needed to improve automated training, which proffers a variety of benefits including ease of deployment, low operational overhead, reduced mass and cost, and an ability to incorporate novel training scenarios as they become relevant. These results inform the design of automated training systems tailored to improve complex skill acquisition in challenging, remote environments, including those of remote medicine, military field operations, and long-duration spaceflight. Substituting operator mediated training with adaptive training capable of attaining similar outcomes is increasingly important with the advent of long-duration human spaceflight and upcoming missions to the moon and Mars.

Future work on automated training should investigate algorithms which actively vary the levels of responsiveness, for instance by adjusting difficulty by amounts proportional to performance, or by shifting staircase progressions based on consecutive positive performance. Moreover, given the surprising benefits of guaranteed challenge in MFP, future work should investigate the use of a minimum difficulty, or floor, during automated training. Further inquiry into the effect of workload on complex, operational task training should be made by investigating self-selection of task difficulty. In addition, a Bayesian approach of predictive dynamic difficulty adjustment may be of significant interest for data-driven training. Finally, future studies should recruit a broad sample of people, including older and non-technical populations, to improve the generality of results.

Data availability statement

The raw data supporting the conclusion of this article will be made available by the authors, without undue reservation.

Ethics statement

The studies involving humans were approved by the University of Colorado Institutional Review Board. The studies were conducted in accordance with the local legislation and institutional

requirements. The participants provided their written informed consent to participate in this study. Written informed consent was obtained from the individual(s) for the publication of any identifiable images or data included in this article.

Author contributions

AV: Conceptualization, Methodology, Project administration, Writing–original draft. EG: Data curation, Formal Analysis, Methodology, Project administration, Writing–original draft. ST: Data curation, Formal Analysis, Methodology, Writing–review and editing. EP: Conceptualization, Methodology, Writing–review and editing. EV: Supervision, Writing–review and editing. TC: Conceptualization, Supervision, Writing–review and editing. AA: Conceptualization, Funding acquisition, Methodology, Supervision, Writing–review and editing.

Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. This research was supported by NASA under Grant No. 80NSSC21K1140.

Acknowledgments

We thank Dr. Sage Sherman, Wyatt Rees, Benjamin Peterson, Caden McVey, Kiah May, and Stella Cross for their assistance in algorithm development, server maintenance, and subject testing. The authors also thank Dr. Stephen Robinson and Dr. Daniel Szafir for their guidance, and the dozens of subjects for their enthusiastic participation in this research study.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

References

Abuhamdeh, S. (2020). Investigating the "flow" experience: key conceptual and operational issues. *Front. Psychol.* 11, 158. doi:10.3389/fpsyg.2020.00158

Agarwal, R., and Karahanna, E. (2000). Time flies when you're having fun: cognitive absorption and beliefs about information technology usage. *MIS Q.* 24 (4), 665. doi:10.2307/3250951

Braun, R. D., and Manning, R. M. (2006). "Mars exploration entry, descent and landing challenges," in 2006 IEEE Aerospace Conference. 2006 IEEE Aerospace Conference, Big Sky, MT, USA (IEEE), 1–18. doi:10.1109/AERO.2006.1655790

Butt, A. L., Kardong-Edgren, S., and Ellertson, A. (2018). Using game-based virtual reality with haptics for skill acquisition. *Clin. Simul. Nurs.* 16, 25–32. doi:10.1016/j.ecns.2017.09.010

Cameirão, M. S., Badia, S. B. i., Oller, E. D., and Verschure, P. F. (2010). Neurorehabilitation using the virtual reality based Rehabilitation Gaming System: methodology, design, psychometrics, usability and validation. *J. NeuroEngineering Rehabilitation* 7 (1), 48. doi:10.1186/1743-0003-7-48

Choi, D. H., Kim, J., and Kim, S. H. (2007). ERP training with a web-based electronic learning system: the flow theory perspective. *Int. J. Human-Computer Stud.* 65 (3), 223–243. doi:10.1016/j.ijhcs.2006.10.002

Constant, T., and Levieux, G. (2019). "Dynamic difficulty adjustment impact on players' confidence," in Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems. CHI '19: CHI Conference on Human Factors in Computing Systems, Glasgow Scotland Uk (Glasgow, Scotland, UK: ACM), 1–12. doi:10.1145/3290605.3300693

Cooper, N., Millela, F., Cant, I., White, M. D., and Meyer, G. (2021). Transfer of training—virtual reality training with augmented multisensory cues improves user experience during training and task performance in the real world. *PLOS ONE* 16 (3), e0248225. doi:10.1371/journal.pone.0248225

Csikszentmihalyi, M., and LeFevre, J. (1989). Optimal experience in work and leisure. *J. Personality Soc. Psychol.* 56 (5), 815–822. doi:10.1037/0022-3514.56.5.815

Dhiman, A., Solanki, D., Bhasin, A., Bhise, A., Das, A., and Lahiri, U. (2016). "Design of adaptive haptic-enabled virtual reality based system for upper limb movement disorders: a Usability Study," in 2016 6th IEEE International Conference on Biomedical Robotics and Biomechatronics (BioRob). 2016 6th IEEE International Conference on Biomedical Robotics and Biomechatronics (BioRob), Singapore, Singapore (IEEE), 1254–1259. doi:10.1109/BIOROB.2016.7523803

Diamond, M. (2022). Will communication delays impact mission controllers? An investigation of mood, performance, and workload during analog missions. MS thesis. Grand Forks, North Dakota, USA: University of North Dakota.

Diener, E., Wirtz, D., Tov, W., Kim-Prieto, C., Choi, D. w., Oishi, S., et al. (2010). New well-being measures: short scales to assess flourishing and positive and negative feelings. *Soc. Indic. Res.* 97 (2), 143–156. doi:10.1007/s11205-009-9493-y

Engeser, S., and Rheinberg, F. (2008). Flow, performance and moderators of challenge-skill balance. *Motivation Emot.* 32 (3), 158–172. doi:10.1007/s11031-008-9102-4

Everson, T., McDermott, C., Kain, A., Fernandez, C., and Horan, B. (2017). Astronaut training using virtual reality in a neutrally buoyant environment. *KnE Eng.* 2 (2), 319. doi:10.18502/keg.v2i2.632

Farr, A., Pietschmann, L., Zürcher, P., and Bohné, T. (2023). Skill retention after desktop and head-mounted-display virtual reality training. *Exp. Results* 4, e2. doi:10.1017/exp.2022.28

Francis, E. R., Bernard, S., Nowak, M. L., Daniel, S., and Bernard, J. A. (2020). Operating room virtual reality immersion improves self-efficacy amongst preclinical physician assistant students. *J. Surg. Educ.* 77 (4), 947–952. doi:10.1016/j.jsurg.2020.02.013

Gabay, Y., Karni, A., and Banai, K. (2017). The perceptual learning of time-compressed speech: a comparison of training protocols with different levels of difficulty. *PLOS ONE* 12 (5), e0176488. doi:10.1371/journal.pone.0176488

Garcia, A. D., Schlueter, J., and Paddock, E. (2020). "Training astronauts using hardware-in-the-loop simulations and virtual reality," in *AIAA scitech 2020 forum. AIAA scitech 2020 forum* (Orlando, FL: American Institute of Aeronautics and Astronautics). doi:10.2514/6.2020-0167

Ghani, J. A., and Deshpande, S. P. (1994). Task characteristics and the experience of optimal flow in human—computer interaction. *J. Psychol.* 128 (4), 381–391. doi:10.1080/00223980.1994.9712742

Golding, J. F. (1998). Motion sickness susceptibility questionnaire revised and its relationship to other forms of sickness. *Brain Res. Bull.* 47 (5), 507–516. doi:10.1016/S0361-9230(98)00091-4

Golding, J. F. (2006). Predicting individual differences in motion sickness susceptibility by questionnaire. *Personality Individ. Differ.* 41 (2), 237–248. doi:10.1016/j.paid.2006.01.012

Gray, R. (2017). Transfer of training from virtual to real baseball batting. *Front. Psychol.* 8, 2183. doi:10.3389/fpsyg.2017.02183

Grimm, F., Naros, G., and Gharabaghi, A. (2016). Closed-loop task difficulty adaptation during virtual reality reach-to-grasp training assisted with an exoskeleton for stroke rehabilitation. *Front. Neurosci.* 10, 518. doi:10.3389/fnins.2016.00518

Guadagnoli, M. A., and Lee, T. D. (2004). Challenge point: a framework for conceptualizing the effects of various practice conditions in motor learning. *J. Mot. Behav.* 36 (2), 212–224. doi:10.3200/JMBR.36.2.212-224

Gupta, S. K., Anand, D. K., Brough, J. E., Schwartz, M., and Kavetsky, R. A. (2008). *Training in virtual environments: a safe, cost-effective, and engaging approach to training*, Maryland, USA: College Park.

Hamari, J., Shernoff, D. J., Rowe, E., Coller, B., Asbell-Clarke, J., and Edwards, T. (2016). Challenging games help students learn: an empirical study on engagement, flow and immersion in game-based learning. *Comput. Hum. Behav.* 54, 170–179. doi:10.1016/j.chb.2015.07.045

Hamblin, C. J. (2005). Transfer of training from virtual reality environments. PhD thesis. Wichita, Kansas, USA: Wichita State University.

Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., et al. (2020). Array programming with NumPy. *Nature* 585 (2020), 357–362. doi:10.1038/s41586-020-2649-2

Homan, D., and Gott, C. (1996). "An integrated EVA/RMS virtual reality simulation, including force feedback for astronaut training," in Flight Simulation Technologies Conference. Flight Simulation Technologies Conference, San Diego,CA,U.S.A. (American Institute of Aeronautics and Astronautics). doi:10.2514/6.1996-3498

Hunicke, R. (2005). "The case for dynamic difficulty adjustment in games," in Proceedings of the 2005 ACM SIGCHI International Conference on Advances in computer entertainment technology. ACE05: International Conference on Advances in Computer Entertainment Technology, Valencia Spain (Valencia, Spain: ACM), 429–433. doi:10.1145/1178477.1178573

Iván Aguilar Reyes, C., Wozniak, D., Ham, A., and Zahabi, M. (2022). An adaptive virtual reality-based training system for pilots. *Proc. Hum. Factors Ergonomics Soc. Annu. Meet.* 66 (1), 1962–1966. doi:10.1177/1071181322661063

Jeelani, I., Han, K., and Albert, A. (2017). "Development of immersive personalized training environment for construction workers," in Computing in Civil Engineering 2017. ASCE International Workshop on Computing in Civil Engineering 2017, Seattle, Washington (American Society of Civil Engineers), 407–415. doi:10.1061/9780784480830.050

Jeun, Y. J., Nam, Y., Lee, S. A., and Park, J. H. (2022). Effects of personalized cognitive training with the machine learning algorithm on neural efficiency in healthy younger adults. *Int. J. Environ. Res. Public Health* 19 (20), 13044. doi:10.3390/ijerph192013044

Jones, E., Oliphant, T., Peterson, P., et al. (2001). SciPy: open-source scientific tools for Python. *Comput. Sci. Eng.* 9, 90.

Kassambara, A. (2023). *Package 'rstatix'. Pipe-friendly framework for basic statistical tests*. version. 0.7.2.

Kenyon, R. V., and Afenya, M. B. (1995). Training in virtual and real environments. *Ann. Biomed. Eng.* 23 (4), 445–455. doi:10.1007/BF02584444

King, S., Boyer, J., Bell, T., and Estapa, A. (2022). An automated virtual reality training system for teacher-student interaction: a randomized controlled trial. *JMIR Serious Games* 10 (4), e41097. doi:10.2196/41097

Kintz, N. M., and Palinkas, L. A. (2016). Communication delays impact behavior and performance aboard the international Space station. *Aerosp. Med. Hum. Perform.* 87 (11), 940–946. doi:10.3357/AMHP.4626.2016

Koenig, A., Novak, D., Omlin, X., Pulfer, M., Perreault, E., Zimmerli, L., et al. (2011). Real-time closed-loop control of cognitive load in neurological patients during robot-assisted gait training. *IEEE Trans. Neural Syst. Rehabilitation Eng.* 19 (4), 453–464. doi:10.1109/TNSRE.2011.2160460

Kumar, D., González, A., Das, A., Dutta, A., Fraisse, P., Hayashibe, M., et al. (2018). Virtual reality-based center of mass-assisted personalized balance training system. *Front. Bioeng. Biotechnol.* 5, 85. doi:10.3389/fbioe.2017.00085

Landon, L. B., Rokholt, C., Slack, K. J., and Pecena, Y. (2017). Selecting astronauts for long-duration exploration missions: considerations for team performance and functioning. *REACH* 5, 33–56. doi:10.1016/j.reach.2017.03.002

Lang, Y., Wei, L., Xu, F., Zhao, Y., and Yu, L-F. (2018). "Synthesizing personalized training programs for improving driving habits via virtual reality," in 2018 IEEE Conference on Virtual Reality and 3D User Interfaces (VR). 2018 IEEE Conference on Virtual Reality and 3D User Interfaces (VR), Reutlingen (IEEE), 297–304. doi:10.1109/VR.2018.8448290

Leek, M. R. (2001). Adaptive procedures in psychophysical research. *Percept. Psychophys.* 63 (8), 1279–1292. doi:10.3758/BF03194543

Levitt, H. (1971). Transformed up-down methods in psychoacoustics. *J. Acoust. Soc. Am.* 49 (2B), 467–477. doi:10.1121/1.1912375

Liu, S.-H., Liao, H.-L., and Peng, C.-J. (2005) Applying the technology acceptance model and flow theory to online e-learning users acceptance behavior, *Issues Inf. Syst.* doi:10.48009/2_iis_2005_175-181

Loftin, R. B., Savely, R. T., Benedetti, R., Culbert, C., Pusch, L., Jones, R., et al. (1997). "Virtual environment technology in training," in *Virtual reality, training's future?* Editors R. J. Seidel and P. R. Chatelier (Boston, MA: Springer US), 93–103. doi:10.1007/978-1-4899-0038-8_11

Love, S. G., and Reagan, M. L. (2013). Delayed voice communication. *Acta Astronaut.* 91, 89–95. doi:10.1016/j.actaastro.2013.05.003

McKinney, W. (2010). "Data structures for statistical computing in python," in Proceedings of the 9th Python in Science Conference, 445, 51–56.

Missura, O. (2015). Dynamic difficulty adjustment. PhD thesis. Bonn: Rheinische Friedrich-Wilhelms-Universität.

Moon, H.-S., and Seo, J. (2020). "Dynamic difficulty adjustment via fast user adaptation," in Adjunct Publication of the 33rd Annual ACM Symposium on User Interface Software and Technology. UIST '20: The 33rd Annual ACM Symposium on User Interface Software and Technology, Virtual Event USA (New York City, New York, USA: ACM), 13–15. doi:10.1145/3379350.3418578

Moskaliuk, J., Bertram, J., and Cress, U. (2013). Training in virtual environments: putting theory into practice. *Ergonomics* 56 (2), 195–204. doi:10.1080/00140139.2012.745623

Ng, Y.-L., Ho, F. K., Ip, P., and Fu, K. w. (2019). Effectiveness of virtual and augmented reality-enhanced exercise on physical activity, psychological outcomes, and physical performance: a systematic review and meta-analysis of randomized controlled trials. *Comput. Hum. Behav.* 99, 278–291. doi:10.1016/j.chb.2019.05.026

Oliveira Dos Santos, W., Bittencourt, I. I., Isotani, S., Dermeval, D., Brandão Marques, L., and Frango Silveira, I. (2018). Flow theory to promote learning in educational systems: is it really relevant? *Rev. Bras. Inform. Educ.* 26 (02), 29. doi:10.5753/rbie.2018.26.02.29

Papakostas, C., Troussas, C., Krouska, A., and Sgouropoulou, C. (2022). Personalization of the learning path within an augmented reality spatial ability training application based on fuzzy weights. *Sensors* 22 (18), 7059. doi:10.3390/s22187059

Park, J., MacRae, H., Musselman, L. J., Rossos, P., Hamstra, S. J., Wolman, S., et al. (2007). Randomized controlled trial of virtual reality simulator training: transfer to live patients. *Am. J. Surg.* 194 (2), 205–211. doi:10.1016/j.amjsurg.2006.11.032

Pinheiro, J., Bates, D., DebRoy, S., Sarkar, D., Heisterkamp, S., Van Willigen, B., et al. (2017). Package 'nlme'. Linear and nonlinear mixed effects models. *version* 3 (1), 274.

Plass, J. L., Homer, B., Pawar, S., Brenner, C., and MacNamara, A. (2019). The effect of adaptive difficulty adjustment on the effectiveness of a game to develop executive function skills for learners of different ages. *Cogn. Dev.* 49, 56–67. doi:10.1016/j.cogdev.2018.11.006

Pollack, I. (1968). Methodological examination of the PEST (parametric estimation by sequential testing) procedure. *Percept. Psychophys.* 3 (4), 285–289. doi:10.3758/BF03212746

Psotka, J. (1995). Immersive training systems: virtual reality and education and training. *Instr. Sci.* 23 (5–6), 405–431. doi:10.1007/BF00896880

Reason, J. T. (1968). Relations Between Motion Sickness Susceptibility, the Spiral After-Effect and Loudness Estimation. *Br. J. Psychol.* 59 (4), 385–393. doi:10.1111/j.2044-8295.1968.tb01153.x

Robertson, J. M., Dias, R. D., Gupta, A., Marshburn, T., Lipsitz, S. R., Pozner, C. N., et al. (2020). Medical event management for future deep Space exploration missions to Mars. *J. Surg. Res.* 246, 305–314. doi:10.1016/j.jss.2019.09.065

Roscoe, A. H., and Ellis, G. A. (1990). *A subjective rating scale for assessing pilot workload in flight: a decade of practical use*. Farnborough, England, UK: Royal Aerospace Establishment Farnborough (United Kingdom), 18.

Rose, F. D., Attree, E. A., Brooks, B. M., Parslow, D. M., and Penn, P. R. (2000). Training in virtual environments: transfer to real world tasks and equivalence to real task training. *Ergonomics* 43 (4), 494–511. doi:10.1080/001401300184378

Russell, J. A., Weiss, A., and Mendelsohn, G. A. (1989). Affect Grid: a single-item scale of pleasure and arousal. *J. Personality Soc. Psychol.* 57 (3), 493–502. doi:10.1037/0022-3514.57.3.493

Saluja, I. S., Williams, D. R., Woodard, D., Kaczorowski, J., Douglas, B., Scarpa, P. J., et al. (2008). Survey of astronaut opinions on medical crewmembers for a mission to Mars. *Acta Astronaut.* 63 (5–6), 586–593. doi:10.1016/j.actaastro.2008.05.002

Sampayo-Vargas, S., Cope, C. J., He, Z., and Byrne, G. J. (2013). The effectiveness of adaptive difficulty adjustments on students' motivation and learning in an educational computer game. *Comput. Educ.* 69, 452–462. doi:10.1016/j.compedu.2013.07.004

Saurav, K., Dash, A., Solanki, D., and Lahiri, U. (2018). "Design of a VR-based upper limb gross motor and fine motor task platform for post-stroke survivors," in 2018 IEEE/ACIS 17th International Conference on Computer and Information Science (ICIS). 2018 IEEE/ACIS 17th International Conference on Computer and Information Science (ICIS), Singapore (IEEE), 252–257. doi:10.1109/ICIS.2018.8466538

Schmidt, R. A. (1975). A schema theory of discrete motor skill learning. *Psychol. Rev.* 82 (4), 225–260. doi:10.1037/h0076770

Sherman, S. O., Jonsen, A., Lewis, Q., Schlittenhart, M., Szafir, D., Clark, T. K., et al. (2023). Training augmentation using additive sensory noise in a lunar rover navigation task. *Front. Neurosci.* 17, 1180314. doi:10.3389/fnins.2023.1180314

Simon, M. A., Toups, L., Howe, A. S., and Wald, S. I. (2015). "Evolvable Mars campaign long duration habitation strategies: architectural approaches to enable human exploration missions," in AIAA SPACE 2015 Conference and Exposition. AIAA SPACE 2015 Conference and Exposition, Pasadena, California (American Institute of Aeronautics and Astronautics). doi:10.2514/6.2015-4514

Sinnott, C., Liu, J., Matera, C., Halow, S., Jones, A., Moroz, M., et al. (2019). "Underwater virtual reality system for neutral buoyancy training: development and evaluation," in 25th ACM Symposium on Virtual Reality Software and Technology.

VRST '19: 25th ACM Symposium on Virtual Reality Software and Technology, Parramatta NSW Australia (Parramatta, New South Wales, Australia: ACM), 1–9. doi:10.1145/3359996.3364272

Siu, K.-C., Best, B. J., Kim, J. W., Oleynikov, D., and Ritter, F. E. (2016). Adaptive virtual reality training to optimize military medical skills acquisition and retention. *Mil. Med.* 181 (5S), 214–220. doi:10.7205/MILMED-D-15-00164

Taylor, M. M., and Creelman, C. D. (1967). PEST: efficient estimates on probability functions. *J. Acoust. Soc. Am.* 41 (4A), 782–787. doi:10.1121/1.1910407

Thurman, R. A., and Mattoon, J. S. (1994). Virtual reality: toward fundamental improvements in simulation-based training. *Educ. Technol.* 34 (8), 56–64.

Trevino, L. K., and Webster, J. (1992). Flow in computer-mediated communication: electronic mail and voice mail evaluation and impacts. *Commun. Res.* 19 (5), 539–573. doi:10.1177/009365092019005001

Vallat, R. (2018). Pingouin: statistics in Python. *J. Open Source Softw.* 3 (31), 1026. doi:10.21105/joss.01026

Vaughan, N., Gabrys, B., and Dubey, V. N. (2016). An overview of self-adaptive technologies within virtual reality training. *Comput. Sci. Rev.* 22, 65–87. doi:10.1016/j.cosrev.2016.09.001

Webster, J., Trevino, L. K., and Ryan, L. (1993). The dimensionality and correlates of flow in human-computer interactions. *Comput. Hum. Behav.* 9 (4), 411–426. doi:10.1016/0747-5632(93)90032-N

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D., François, R., et al. (2019). Welcome to the tidyverse. *J. Open Source Softw.* 4 (43), 1686. doi:10.21105/joss.01686

Wilbert, J., Grosche, M., and Gerdes, H. (2010). Effects of evaluative feedback on rate of learning and task motivation: an analogue experiment. *Learn. Disabil. A Contemp. J.* 8 (2), 43–52.

Xue, S., Wu, M., Kolen, J. F, Aghdaie, N., and Zaman, K. A. (2017). "Dynamic difficulty adjustment for maximized engagement in digital games," in Proceedings of the 26th International Conference on World Wide Web Companion - WWW '17 Companion. the 26th International Conference, Perth, Australia (Perth, Western Australia, Australia: ACM Press), 465–471. doi:10.1145/3041021.3054170

Yang, X., Wang, D., and Zhang, Y. (2016). "An adaptive strategy for an immersive visuo-haptic attention training game," in *Haptics: perception, devices, control, and applications*. Editors F. Bello, H. Kajimoto, and Y. Visell (Cham: Springer International Publishing (Lecture Notes in Computer Science), 441–451. doi:10.1007/978-3-319-42321-0_41

Yovanoff, M., Pepley, D., Mirkin, K., Moore, J., Han, D., and Miller, S. (2017). "Personalized learning in medical education: designing a user interface for a dynamic haptic robotic trainer for central venous catheterization," in Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 615–619. doi:10.1177/1541931213601639

Zohaib, M. (2018). Dynamic difficulty adjustment (dda) in computer games: a review. *Adv. Human-Computer Interact.* 2018, 1–12. doi:10.1155/2018/5681652

Zuzula, E. A., Dixon, J. B, Davis, E., Bretl, K., Pinedo, C., and Clark, T. K. (2018). "A numerical algorithm to estimate an achievability limit for crewed planetary landing," in 2018 IEEE Aerospace Conference. 2018 IEEE Aerospace Conference, Big Sky, MT (IEEE), 1–7. doi:10.1109/AERO.2018.8396484