Automated Assessment and Adaptive Multimodal Formative Feedback Improves Psychomotor Skills Training Outcomes in Quadrotor Teleoperation

Emily Jensen
emily.jensen@colorado.edu
University of Colorado Boulder
Boulder, Colorado, USA

Sriram Sankaranarayanan
srirams@colorado.edu
University of Colorado Boulder
Boulder, Colorado, USA

Bradley Hayes
bradley.hayes@colorado.edu
University of Colorado Boulder
Boulder, Colorado, USA

ABSTRACT

The workforce will need to continually upskill in order to meet the evolving demands of industry, especially working with robotic and autonomous systems. Current training methods are not scalable and do not adapt to the skills that learners already possess. In this work, we develop a system that automatically assesses learner skill in a quadrotor teleoperation task using temporal logic task specifications. This assessment is used to generate multimodal feedback based on the principles of effective formative feedback. Participants perceived the feedback positively. Those receiving formative feedback viewed the feedback as more actionable compared to receiving summary statistics. Participants in the multimodal feedback condition were more likely to achieve a safe landing and increased their safe landings more over the experiment compared to other feedback conditions. Finally, we identify themes to improve adaptive feedback and discuss and how training for complex psychomotor tasks can be integrated with learning theories.

CCS CONCEPTS

• **Human-centered computing** → *Empirical studies in HCI*; • **Theory of computation** → *Modal and temporal logics*; • **Computing methodologies** → *Artificial intelligence.*

KEYWORDS

Formative Feedback, Automated Assessment, Training

ACM Reference Format:

Emily Jensen, Sriram Sankaranarayanan, and Bradley Hayes. 2024. Automated Assessment and Adaptive Multimodal Formative Feedback Improves Psychomotor Skills Training Outcomes in Quadrotor Teleoperation. In *Proceedings of Make sure to enter the correct conference title from your rights confirmation email (HAI '24).* ACM, New York, NY, USA, 10 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

Future industrial development will depend on collaboration between humans and automated systems. While some fear losing jobs to automation, experts argue there will be a need for highly-skilled human-automation teams that can adapt to customer-specific tasks [3, 27]. Humans in these collaborative teams must be able to understand how the autonomous system works, how to manage it, and how to adapt when maintenance is needed or other technical issues arise [11]. A recent report estimates that one third of job requirements will require technological skills that are not yet considered crucial [40], meaning that employees will need to continually adapt as technological innovation continues.

Some of the most significant barriers to achieving this industrial development are the ability to scale up capacity as well as upskill and reskill the current workforce [28]. Experts estimate that 50% of existing employees will need to be retrained or upskilled by 2025 to keep up with technological advancement, placing significant pressure on both employers and employees to meet these demands [21, 40]. With the recent developments of artificial intelligence capabilities, researchers are considering how to improve and automate this crucial training.

Training systems and programs for developing industrial skills are a promising opportunity to expand the workforce. For example, sub-baccalaureate training programs and stackable certifications may allow disadvantaged workers to access the training needed to enter highly-skilled industrial sectors [1]. In order to achieve this goal, training programs will need to focus on transferable skills and present interfaces that are "customizable, individualized, and on-demand" to address the needs of each unique learner [15].

Current training methods such as individualized instruction and pre-recorded modules cannot scale up to meet this need to upskill. They also ignore the fact that many employees enter training with skills that can be transferred to a new task. Intelligent Tutoring Systems are designed to meet just these demands in classrooms by developing personalized models of students and building on knowledge the student has already mastered. Although previous work discusses applying these approaches outside the classroom [39], existing approaches for training physical tasks has not been systematically researched and integrated with learning theory.

In this work, we develop a system that provides formative feedback on a quadrotor landing task using automated assessment from temporal logic task specifications and generative artificial intelligence. We demonstrate that foundational research in feedback can be transferred from school-based educational technology to develop adaptive training systems for complex, multi-objective interaction tasks between humans and autonomous systems.

HAI '24, June 03–05, 2018, Woodstock, NY

ACM ISBN 978-1-4503-XXXX-X/18/06
https://doi.org/XXXXXXX.XXXXXXX

2 RELATED WORK

This work focuses on training for complex psychomotor tasks. We define complex tasks as those with high variability, requiring multiple steps to complete, and whose performance depends on multiple factors [45]. Psychomotor tasks require the coordination of physical (grasping, teleoperating) and cognitive (planning, decision making) elements to successfully complete the task [29].

2.1 Automated Assessment for Complex Psychomotor Tasks

Adaptive training systems must be able to automatically assess performance before providing feedback. This is especially difficult for complex psychomotor tasks because successful performance depends on a variety of factors.

Assessment is also highly dependent on the task domain; as such, previous work in automated assessment has developed specialized methods for the specific domain. For example, Rauter et al. analyzed performance on a rowing task by comparing the velocity profile of the rowing stroke against expert performance [37]. Other studies similarly compare performance to expert trajectories as a benchmark for successful performance [6, 41]. Surgical robotics studies have used physiological metrics such as smoothness, motion amplitudes, and muscular activation [44] in addition to response time for unanticipated events [46]. A recent study evaluated performance in human-robot teaming using number of collisions, number of re-grasps, and total task time [35]. The metrics presented here are largely *outcome-based*, meaning they provide an overall indication of task success, but lack a nuanced description of the learner's *process* of completing the task.

Additionally, the methods used to assess performance are not scalable to task variants or new domains. Recent examples of automated assessment (or more simply, error detection) include domains such as table tennis [24], martial arts [9, 10, 33], piano [26], medical first responders [34], industrial production [12], and surgery [38]. Many of these methods rely on neural network classifiers, which require significant data to train and do not provide explanations for their predictions. In this work, we build off Jensen et al.'s proposed framework, which describes skill as a vector of several performance outcomes [16]. They assess performance relative to a set of task specifications, which are easy to compute and require no data to learn. Task specifications use logical requirements and constraints to define task objectives, such as reaching the goal state within T time: $\textsc{Eventually}_{[0,T]}\, atGoal$.

2.2 Formative Feedback

Providing feedback is key to improving a learner's performance. Summative feedback provides a general summary of performance after the learning program is completed [43]. While useful for providing an overview of performance, learners are left to self-regulate their practice in the absence of other feedback. On the other hand, formative feedback is provided during the learning process to help guide future learning [42]. This type of feedback is given more frequently and focuses on encouragement. Based on recent reviews in the educational technology literature, we identified the following elements of effective formative feedback:

- *Reflection*: feedback gives detailed information about the task, process, and encourages the learner to self-reflect [2, 14, 23, 32].
- *Motivation*: feedback expresses confidence in the learner's abilities [13, 23, 32].
- *Timely*: feedback is directly connected to the learner's recent actions [13, 23].
- *Actionable*: feedback provides specific guidance for improvement that is related to the assessment criteria [13, 14, 23].
- *Manageable*: feedback is detailed but not overwhelming to interpret [13, 14].

These elements of feedback have been shown to support learning outcomes and are positively perceived by students in classroom learning settings. One goal of this study is to evaluate whether this theory of effective feedback improves task performance in a complex psychomotor task domain.

2.3 Training for Psychomotor Tasks

Recent work in developing end-to-end adaptive training systems for complex psychomotor tasks has focused on individual domains such as surgery, sports, marksmanship, karate, driving, aircraft maintenance, and additive manufacturing [19, 39, 49].

Several works have discussed how pedagogically-informed feedback strategies may be implemented in training systems. For example, Korhonen et al. [18] and Pérez-Ramírez et al. [36] discuss how theories such as embodied cognition can be implemented into virtual reality learning environments. Other work proposes inserting erroneous solutions to encourage critical thinking [4] or using adaptive epistemic feedback for training [22]. However, none of these studies have implemented and evaluated the effectiveness of these theories.

Training systems that provide performance feedback tend to rely on prerecorded responses or templates for reacting to failure modes [6, 7, 31] or display statistical summaries of key performance outcomes [37, 41, 49]. An ultrasound placement study generated a visual comparison between the learner's placement and orientation compared to an expert [41]. To the best of our knowledge, no studies have investigated the use of generated natural-language text to provide formative feedback to learners. In this study, we provide statistical summaries as a baseline condition and compare learner performance to generated text containing the identified elements of effective formative feedback.

3 CONTRIBUTIONS AND RESEARCH QUESTIONS

The contribution of this paper is **a flexible and validated framework for automatically assessing performance in multi-objective tasks and generating personalized formative feedback**. We accomplish this by pairing robustness measures of formal task specifications with natural language feedback generated from pedagogically-grounded templates. In this study, we compare groups that received summary metrics of their performance, automatically generated text feedback, and text feedback paired with an annotated figure showing their trajectory. We evaluate the system using the following research questions (RQs).

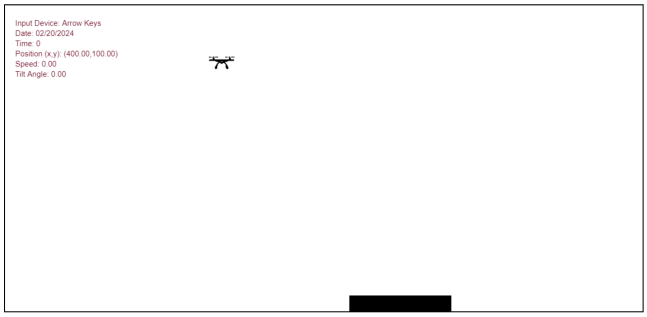


Figure 1: Screenshot of the starting configuration of the quadrotor landing task.

- *RQ 1.* Do participants perceive the elements of formative feedback differently between score-based, semantic, or multimodal presentations?
- *RQ 2.* What factors predict the perception of formative feedback?
- *RQ 3.* How does automated formative feedback affect participants' learning trajectory?

4 METHODS

The study code and analysis scripts will be added as a link here in the final version of the paper.

4.1 Quadrotor Landing Task

In this experiment, participants completed a simulated quadrotor landing task. In this task, participants used keyboard inputs to adjust the quadrotor's throttle (vertical force) and attitude (rotation for horizontal force). To achieve a safe landing, the quadrotor must reach the landing pad with a speed less than 15 m/s and a rotation angle within ±5°. We labeled a landing attempt as unsafe if the drone reached the landing pad but did not satisfy the required speed or angle constraints. All other landing attempts were crashes. We refer to a landing attempt as a *trial*. Each trial was capped at 120 seconds. Figure 1 shows the initial configuration of the task participants completed. The initial position of the drone and the landing pad did not change between trials. Yuh et al.'s work provides more details on the dynamics of the quadrotor and design of the simulator [47].

4.2 Participants

We recruited participants using the Prolific platform. All participants were United States residents. 177 participants completed the study. Of these participants, 16 restarted the study due to technical issues. To minimize confounding learning effects, we excluded six participants that completed more than five trials before restarting. We excluded another four participants who did not provide a good faith effort in the experiment, as measured by never using the horizontal input controls and crashing the quadrotor on each trial. This resulted in a final dataset of 167 participants.

Participants ranged in age from 18 to 74 years, with a median age of 35 years. Their reported gender identities were 73 Men (44%), 87 Women (52%) and 7 Non-binary individuals (4%). 97% of participants reported no prior experience flying drones or have flown a drone a few times. Participants reported a range of video game experience, with 30 not playing video games (18%), 46 playing monthly (27.5%), 40 playing weekly (24%) and 51 playing daily (30.5%).

4.3 Experiment Design and Procedure

We conducted a between-subjects study with three experimental conditions. In the baseline condition, participants received summary statistics such as their landing outcome and an overall score of their performance, replicating prior work [48]. In the second condition, participants received AI-generated text feedback, described in Section 4.5. In the third condition, participants received AI-generated text feedback along with an annotated image of their trajectory, which highlighted an area of their trajectory to focus on improving. In our final dataset, 55 participants were in the baseline condition, 56 participants were in the text feedback condition, and 56 participants were in the multimodal feedback condition.

After consenting to participate in the study and reading the instructions for the task, participants completed the quadrotor landing task. After each trial, participants received feedback on their performance depending on their experimental condition. Participants then rated the feedback they received and completed the landing task again. After completing the task 20 times, participants completed a brief demographic questionnaire and rated their overall perception of the feedback they received. On average, participants completed the experiment in 29.12 minutes (SD = 10.24 minutes). They spent an average of 28.12 seconds (SD = 11.08 seconds) reviewing and rating their feedback on each trial.

4.4 Automated Assessment

The system automatically assessed landing performance using a previously validated framework [16]. For each component of the task, we defined a specification using signal temporal logic [8], a formalism for specifying complex temporal tasks. For the quadrotor landing task, the specifications focused on the safety and landing behaviors. The specifications are shown in Table 1. Robustness values are a quantitative score that describes how well the trajectory of the quadrotor meets the given specification; large positive values indicate better compliance (e.g., staying far away from the edge of the simulation window) while large negative values indicate stronger violations (e.g., extreme landing angle) [8, 16].

To keep the feedback manageable, we used a heuristic for selecting the top area of improvement the participant should focus on for the next trial. The safety components were given the highest priority; if the quadrotor crashed into any of the sides of the simulation window (indicated by $s_i = 0$), this was selected as the area of improvement. If the quadrotor landed unsafely, either landing speed or angle was chosen as the area of improvement (l_3 or $l_4 < 0$). For successful landings, the area of improvement was selected as overall efficiency if the trial was longer than a predetermined length or otherwise defaulted to smoothness.

4.5 Formative Feedback Design

Participants received formative feedback based on the context generated from the automated assessment in Section 4.4 and natural language generated from a prompt incorporating the elements of

Table 1: Overview of specifications for quadrotor landing task with range of possible robustness values for the individual components. Note that the specific values for s_i and l_i depend on the size of the simulation window and the quadrotor.

Description	Specification	Robustness Range
Avoid left edge	$s_1 = x > 0$	[0, 1210]
Avoid right edge	$s_2 = x < 1250$	[0, 1210]
Avoid bottom edge	$s_3 = y > 0$	[0, 575]
Avoid top edge	$s_4 = y < 600$	[0, 575]
Avoid left land edge	$l_1 = x > 650$	[-650, 560]
Avoid right land edge	$l_2 = x < 850$	[-360, 850]
Slow landing speed	$l_3 = \|\|v\|\| < 15$	[-17, 15]
Shallow landing angle	$l_4 = \|\phi\| < 5$	[-24, 5]
Safety component	$S = \wedge_{i=1}^{4} s_i$	
Landing component	$L = \wedge_{i=1}^{4} l_i$	
Complete task in 120s	S UNTIL$_{[0,120]}$ L	

formative feedback discussed in Section 2.2. The prompt included a description of the target task, the identified area of improvement, the generated image of the trajectory, and an explanation of what each element of the feedback should contain. We used GPT-4V [30] to generate the text feedback.

The visual feedback consisted of an image of the landing trajectory with a superimposed circle to highlight a specific area of improvement along the trajectory. We identified the location of the circle using the area of improvement heuristic described in Section 4.4. In the event of a crash, we placed the circle on the location where the quadrotor crashed. If the quadrotor landed unsafely, we placed the circle at the point in the last 50 steps in the trajectory that had the worst robustness for landing speed or landing angle. For a safe landing, we placed the circle at the point in the trajectory with the highest combined control inputs.

Figure 2 shows an example of each type of feedback. We generated the full set of text and image feedback regardless of condition so participants waited the same amount of time between trials.

4.6 Measures

Subjective Measures. After each trial, participants rated the feedback they received. The purpose of the survey items was to understand how the generated feedback aligned with the desired dimensions of formative feedback described in Section 2.2. Table 2 summarizes the survey items participants completed after each trial. After completing the experiment, participants completed an exit survey that recorded their gender identity, age, experience flying drones, and video game experience. Participants also rated how helpful the feedback was overall ("The feedback I received helped me perform better on the task"; 1 = Strongly disagree, 5 = Strongly agree) and provided a text response discussing how the feedback influenced their piloting strategy over time.

Objective Measures. We recorded the trajectory for each trial. The trajectory data included the quadrotor's x and y position and velocity, the quadrotor's rotation, and the participant's control throttle and attitude inputs. For each time step, we also calculated the trajectory's robustness according to the specifications in Section 4.4. We recorded both trajectory and robustness data at 50 Hz.

4.7 Data Analysis

RQs 1 and 2 ask how participants perceived the feedback they received. The variables of interest were the subjective measures on each feedback dimension shown in Table 2 and the overall rating of feedback helpfulness, which yielded ordinal values. We found little discrimination between the extreme values of the Likert scales (Strongly Agree vs. Agree and Strongly Disagree vs. Disagree) so we collapsed these measures to a three-point scale (Disagree, Neutral, Agree) for analysis.

To answer RQ 1, we used the Kruskal-Wallis H-test to test for differences in feedback ratings between groups. As mentioned above, participants rated each dimension of feedback after every trial. To create independent samples, we aggregated survey responses for each participant across trials by calculating the most common response for each item. We found that participant ratings do not change much over time, which suggests that this method of aggregation provides an overall rating of each dimension of feedback.

We used ordered logistic regression models to answer RQ 2. The outcome variables were each participant's overall rating for each feedback dimension and their overall rating of the feedback's helpfulness. The independent variables included participant demographics, total number of safe landings, average trial time, and average time spent reviewing feedback. We also performed a trial-wise analysis of the feedback ratings, using trial time, feedback time, trial number, and landing outcome as predictors. The coefficients of these models (β) represent log-odds; we also report odds-ratios as OR to aid with interpretation.

For RQ3, we considered several metrics of learning trajectory. We first evaluated mastery of the quadrotor landing task by calculating how many participants in each condition achieved at least one safe landing across the 20 trials. We used Fisher's Exact Test to test for differences between feedback conditions. We also considered how much participants improved in the task over time. We measured this by calculating how many more safe landings each participant achieved in the second half of the trials compared to the first half. To compare differences between feedback conditions, we used an independent-samples t-test.

5 RESULTS

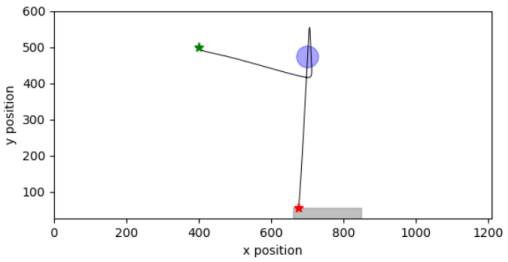
RQ 1 asks if participants in different feedback conditions perceived the dimensions of formative feedback differently. There is a statistically significant difference in ratings along the manageable ($H(2) = 18.0,\ p < 0.001$) and actionable ($H(2) = 18.1,\ p < 0.001$) dimensions. Post-hoc Dunn's test with Bonferroni corrections reveals a significant difference in ratings along the manageable dimension between the baseline and text feedback conditions ($p < 0.001$) and between the baseline and multimodal feedback conditions ($p = 0.005$). There is also a significant difference in ratings along the actionable dimension between the baseline and text feedback conditions ($p < 0.001$) and between the baseline and multimodal feedback conditions ($p < 0.001$). There are no differences in ratings

Trial No.	Total Time (seconds)	Final Velocity (m/s)	Final Angle (°)	Landing	Score
1	4.22	21.4	9.724	Unsafe	552.6

(a) Baseline feedback condition

Firstly, you've done a good job at keeping the drone within the safe flight area throughout your flight path, which is crucial for avoiding any possible incidents. However, the area you can most improve in is the landing speed. Specifically, your approach speed on landing needs to be more controlled; it's essential to ensure the drone is not moving too quickly as it touches down on the landing pad. To improve on this, focus on reducing thrust a little earlier as you approach the landing area, allowing the drone to decelerate more smoothly. As you reflect on your flight, consider at what point during your descent you could start easing off the thrust to achieve a gentler landing. I am confident that with this adjustment, you'll master the precision required for a smooth and safe touchdown, enhancing your piloting skills significantly. Keep up the determined effort; you're on the right track to completing the target task with excellence.

(b) Text feedback condition



Great job maintaining the drone within the safe flying zone throughout your trajectory; you've shown that you're capable of effectively controlling the drone within the set boundaries. However, there's a significant opportunity for improvement in the way you manage your landing speed. For your next attempt, focus on gradually reducing thrust as you approach the landing pad to ensure the drone descends gently and its speed is well below the target upon touchdown. Reflect on how you can initiate the descent earlier and with less abrupt changes in altitude, which will help you achieve a smoother landing. I'm confident in your ability to adapt and apply these strategies, as the control you've already demonstrated suggests that with a little fine-tuning, you'll be able to nail the perfect landing speed. Keep up the good work and continue to practice these adjustments for an even better performance next time.

The trajectory above has been highlighted to show which part of your performance you can focus on improving.

(c) Multimodal feedback condition

Figure 2: Examples of the three feedback conditions used in the experiment.

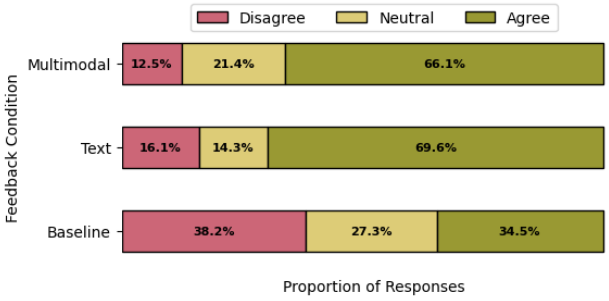
between the non-baseline feedback conditions. There are no significant differences in ratings for the motivation, timely, or reflection dimensions between the feedback groups.

A closer investigation of the distributions of survey responses shows that while participants in all conditions rate the manageability of the feedback as the right amount of information (50-60% of participants in each condition), more participants rate the generated feedback conditions as providing too much information (32-41%). Participants in the baseline condition are more likely to rate the feedback as having not enough information (29%). More participants receiving generated feedback agree that the feedback was actionable (66-70% of participants), while only 35% of participants in the baseline condition agree that the feedback was actionable. Figure 3 shows the distributions of ratings for the manageable and actionable feedback dimensions.

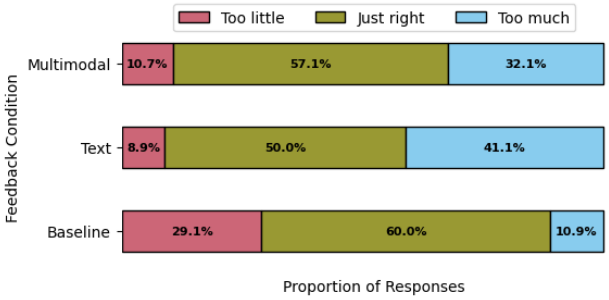
Although there are no significant differences between groups, participants as a whole generally find the feedback to be motivational (58-64% agree) and prompting reflection (64-70% agree).

Table 2: Summary of the survey items participants completed after receiving feedback for a given trial. Participants rated their feedback using these items for each of the 20 trials in the experiment.

Feedback Dimension	Survey Item	Response Options
Motivation	"The feedback motivated me to do better in future trials."	1 = Strongly disagree, 5 = Strongly agree
Manageable	"How much information did the feedback give?"	1 = Much too little, 5 = Much too much
Actionable	"The feedback suggestions were actionable."	1 = Strongly disagree, 5 = Strongly agree
Timely	"How often was the feedback presented?"	1 = Much too infrequent, 5 = Much too often
Reflection	"The feedback prompted me to reflect on my performance."	1 = Strongly disagree, 5 = Strongly agree



(a) Distribution of responses for actionable feedback dimension: "The feedback suggestions were actionable."



(b) Distribution of responses for manageable feedback dimension: "How much information did the feedback give?"

Figure 3: Distributions of responses for feedback dimensions that were significantly different between groups. We show the responses collapsed to a three-point likert scale.

Participants also report similar ratings for the timely dimension, with 25-39% reporting that feedback was delivered too often.

There is no significant difference between groups regarding where they found the feedback helpful to improving their performance. The majority of participants in the generated feedback conditions agree that the feedback benefited their performance (61-66%) and 45% of participants in the baseline condition agree that the feedback helped their performance on the task.

RQ 2 asked what factors predict the perception of formative feedback. We first fit an ordered logistic regression model to predict the rating of whether the feedback helped improve performance. While none of the demographic variables are significant predictors, four of the five aggregate measures of formative feedback are significant ($p < 0.05$). Participants with higher 'motivation' ($\beta = +0.97,\ OR = 2.36$) and higher 'reflection' ($\beta = +0.76,\ OR = 2.14$) responses are more likely to rate the feedback as more helpful. Participants with higher 'timely' responses are more likely to rate the feedback as less helpful ($\beta = -1.12,\ OR = 0.33$), since higher 'timely' ratings correspond to the perception that the feedback was given too often. Those with higher 'manageable' ratings rate the feedback as more helpful ($\beta = +1.10,\ OR = 3.01$), which is interesting because higher manageable ratings mean the feedback contained too much information.

There are few variables that predict overall ratings for the elements of formative feedback. Participants with more experience with drones were more likely to rate the manageable dimension as having not enough information ($\beta = -0.80,\ OR = 0.45$). Those who achieved more safe landings are more likely to rate the feedback as occurring too often ($\beta = +0.09,\ OR = 1.10$). Older participants rate the feedback as promoting more reflection ($\beta = +0.06,\ OR = 1.06$). Finally, participants that spent more time completing the experiment rate the feedback as promoting less reflection, although this difference is small ($\beta = -0.001,\ OR = 1.00$).

Table 3 summarizes which variables predicted feedback ratings at the trial level. Motivation, reflection, and actionable ratings increase with longer trial times and more time spent reviewing feedback. Participants view feedback more negatively as the experiment progresses, with higher trial numbers corresponding to lower motivation, reflection, and actionable ratings. Participants also rate the feedback as containing too much information and occurring too often when they achieve more successful landings.

RQ 3 asks how learning trajectories differed between groups. Fisher's Exact Test shows a significant difference in the number of people who failed to achieve a safe landing in any of the trials between the multimodal feedback condition and the baseline condition ($p = 0.04$). Only two participants fail to achieve a safe landing in the multimodal feedback condition (3.6%), compared to eight participants in the baseline condition (14.5%). There are no differences between the multimodal and text feedback conditions or the text and baseline feedback conditions.

All groups improved their performance at the task, as demonstrated by fewer crashes and more safe landings in the second half of the trials (see Table 4). Participants in the multimodal feedback condition show a larger increase in safe landings ($M = 2.4$ more safe landings, $SD = 2.3$) compared to the text feedback condition ($M = 1.4$, $SD = 2.7$), $t(110) = 2.2$, $p = 0.03$. There are no significant

Table 3: Predictors of trial-wise feedback ratings for each dimension. We report only significant ($p < 0.05$) coefficients as β with corresponding odds-ratios.

Predictor	Motivation (β, OR)	Manageable (β, OR)	Timely (β, OR)	Reflection (β, OR)	Actionable (β, OR)
Trial Time (s)	+0.005, 1.00		-0.01, 0.99	+0.01, 1.01	+0.004, 1.00
Feedback Time (s)	+0.01, 1.01		+0.01, 1.01	+0.01, 1.01	+0.01, 1.01
Trial Number	-0.03, 0.97		+0.02, 1.02	-0.03, 0.97	-0.02, 0.98
Type of Landing		+0.15, 1.17	+0.33, 1.40		

differences between the baseline condition and the other feedback conditions.

6 DISCUSSION

6.1 Main Findings

In this paper, we developed an end-to-end training system that assesses performance and provides actionable feedback on a quadrotor landing task with no human intervention. The system uses temporal logic task specifications and task demonstrations to assess performance and provide formative feedback to the learner. We found significant differences in how manageable and actionable participants in the different conditions perceived the feedback. Importantly, we found several differences in learning outcomes between conditions. **Participants receiving multimodal feedback were more likely to safely land the quadrotor and showed greater improvement in safe landings in the second half of trials** compared to other feedback conditions.

Overall, participants in all conditions had favorable views of the feedback they received. In particular, participants in the formative feedback conditions mentioned how the feedback impacted their motivation and self-confidence. One participant in the text feedback condition noted, "Overall the encouragement was genuinely nice to receive, and helped to give motivation in completing the task and wanting to do well." Participants in the text and multimodal feedback conditions also reported the feedback felt personalized to their own skills and struggles. Another participant receiving text feedback reported, "The feedback actually felt tailored to me, and not just the same stock answer every time."

Participants in the baseline condition showed surprisingly positive perceptions of their feedback. The written responses indicated that participants were motivated by wanting to figure out how to improve their performance score. One participant noted, "I tried to tell which criteria affected the score more, and how." Although we did not specifically design the baseline condition to be motivational and engaging, this result is in line with work showing that feedback can be intrinsic to the learner [5]. Future works can investigate how to integrate feedback with principles of self-regulated and gamified learning.

However, we found that participants naturally differentiated between performance data and formative feedback. In particular, participants in the baseline condition pushed back against labelling the data summary as feedback. Participants in this condition noted, "The feedback did not help much with strategizing, but it did make me want to get better scores." and "The feedback didn't seem like feedback, because there was no suggestions. The feedback ... was just the numbers that we scored." These findings show that both the content and the delivery of feedback matters. Many of the studies discussed in Section 2 implemented feedback similar to our baseline condition in the form of summary metrics. This feedback may be effective by providing learners with more information about their performance, but this depends on the learner to be able to interpret and devise new strategies based on their data. Truly formative feedback should help the learner interpret their data and act on it in future practice.

Participants receiving multimodal feedback were more likely to achieve a safe landing compared to the baseline condition. This could be due to how the multimodal feedback was personalized to address the learner's greatest area of improvement, while the summary statistics in the baseline condition remained the same regardless of performance. For example, one participant in the multimodal feedback condition said, "The feedback helped tremendously by showing the exact location the unnecessary movements where at." Additionally, the format of the formative feedback may have encouraged participants to experiment with new control strategies. Another participant noted, "The feedback helped me feel more confident in the adjustments I was making and to try new approaches."

Finally, participants in the multimodal feedback condition improved more than the text feedback condition by achieving more safe landings in the second half of the experiment compared to the first half. This may be due to additional information provided by the annotated trajectory in the multimodal feedback condition. With the annotated trajectory, participants can pair general ideas presented in the text with a concrete emphasis on a particular area highlighted on the trajectory. In both conditions, participants noted that the feedback did not give specific enough strategies to improve performance. One participant in the text feedback condition noted, "It asked me to consider how changing it "might" be more effective but not exactly how (try using the W key more often to keep the drone up longer, for example)."

These observations highlight an area to improve the feedback prompt template. When designing the feedback, we prompted the model to use actionable suggestions related to the throttle and rotation of the quadrotor. While these terms are specific to quadrotors and other aircraft, they did not tell the learner exactly what to do (e.g. what buttons to push and how) to perform better in this particular simulation environment. This suggests feedback can be actionable on several levels, depending on the complexity of the task one is learning.

Table 4: Average (SD) number of landings for each feedback condition, calculated for the first and second half of the trials and for the whole experiment. Improvement is the average (SD) number more landings in the second half of trials.

	Safe Landings				Safe and Unsafe Landings			
Condition	Trials 1-10	Trials 11-20	Improvement	All Trials	Trials 1-10	Trials 11-20	Improvement	All Trials
Baseline	3.42 (2.85)	5.40 (3.28)	1.98 (2.18)	8.82 (5.75)	7.04 (2.24)	8.53 (1.93)	1.49 (1.92)	15.56 (3.71)
Text	3.05 (2.96)	4.41 (3.07)	1.36 (2.65)	7.46 (5.42)	6.55 (2.61)	8.20 (2.02)	1.64 (2.19)	14.75 (4.17)
Multimodal	2.75 (2.46)	5.12 (2.52)	2.38 (2.33)	7.88 (4.39)	6.93 (2.21)	8.89 (1.27)	1.96 (1.93)	15.82 (3.06)

6.2 Emerging Themes

The results from this study illuminate a need to consider how to adapt feedback beyond the most recent trial. For example, the approach presented here does not consider persistent skill gaps that appear over several trials. One participant in the text feedback condition wrote, "I wish the feedback generated built on the performance in prior trials so the feedback could say *you've improved!* instead of *you need to be better at the same thing... even though you actually did improve compared to the last trial.*"

Additionally, we can use different feedback strategies depending on the overall task performance; high-performing individuals may only need to reinforce their successful control strategies while novices may need more structured and specific feedback presented in this study. Several participants noted frustration when receiving feedback after a successful landing. One participant in the text feedback condition noted, "It is a little discouraging to finally make a successful landing, and then get *a yeah, you did it - but you should focus on doing it better.*"

Future work should also consider how to schedule feedback over time. Several participants reported ignoring the feedback as the trial progressed, especially if they were consistently performing well on the task. A participant receiving multimodal feedback wrote, "After finding the fastest way of landing the drone, I did not follow any more suggestions." Additionally, Participants also reported needing time to independently explore the dynamics of the task before receiving performance feedback. A participant in the text feedback condition reported, "It would benefit me to go straight into the next trial so I can continue to make small adjustments back to back... Half of learning is trial and error."

Recent work discusses how prompt-based generative feedback methods are ideal for quickly prototyping and testing feedback templates [17]. Future works can investigate using simple rules to determine what feedback template to generate. How to adjust the timing of formative feedback based on the number of attempts and performance outcomes remains an open question.

As automated training systems continue to develop, it is important to consider their place among other workplace programs. It is likely we will need to balance automated approaches with more traditional one-on-one training [15]; in addition to learning technical skills, training programs will need to consider the social aspects of learning such as developing a community of practice within an organization [20]. As required workplace skills and knowledge continue to develop over time, training systems will need to both provide initial background knowledge and additional support to help workers remain up-to-date [25].

6.3 Study Limitations

This work is limited in several ways. First, the quadrotor landing task did not change between trials. This means that when participants found a control strategy that yielded a successful result, they tended to repeat the same strategy. Future works may wish to randomize the starting point of the drone in the simulation to provide more insight about if participants are learning strategies that transfer to other landing scenarios.

The other main limitation of this study is the online nature of the data collection. While the Prolific platform allowed us to quickly recruit a large sample of participants, we were not able to observe nuanced reactions to the feedback they received. Future work can pair crowd-sourced methods with in-person studies to understand how participants choose to integrate feedback into their learning process.

7 CONCLUSIONS

In this paper, we developed an adaptive training system for a simulated quadrotor landing task. The system first assesses performance based on temporal logic specifications, which require no prior data to learn and can be flexibly adapted to new tasks and situations. Using these assessment results, we automatically generated multimodal feedback adhering to principles of effective formative feedback. While participants in all conditions reported finding the feedback engaging and motivating, they differed in their ratings of how actionable and manageable the feedback was. Since the goal of a training system is to help learners master a new task, we also considered learning differences between conditions. We found that participants receiving multimodal feedback were more likely to achieve a safe landing. They also improved more over the course of the experiment by increasing their safe landings more compared to other feedback conditions. Based on these results, we identified future opportunities to further adapt feedback over time and consider the learner's affective state when delivering feedback. While future work in psychomotor task training will continue to depend on domain-specific methods and knowledge, we encourage researchers to align their methods with established pedagogical theories of learning and feedback.

ACKNOWLEDGMENTS

This work was supported by the US National Science Foundation (NSF) under award number 1836900.

REFERENCES

[1] Megan Andrew, Timothy Marler, Jesse Lastunen, Hannah Acheson-Field, and Steven Popper. 2020. *An Analysis of Education and Training Programs in Advanced*

Manufacturing Using Robotics. RAND Corporation, Pittsburgh, Pennsylvania. https://doi.org/10.7249/RR4244

[2] Seyyed Kazem Banihashem, Omid Noroozi, Stan Van Ginkel, Leah P. Macfadyen, and Harm J.A. Biemans. 2022. A systematic review of the role of learning analytics in enhancing feedback practices in higher education. *Educational Research Review* 37 (Nov. 2022), 100489. https://doi.org/10.1016/j.edurev.2022.100489

[3] D. A. Bell. 1985. Employment skills for the robot age. *Robotica* 3, 2 (April 1985), 93–95. https://doi.org/10.1017/S0263574700001788

[4] C. Buche, R. Querrec, P. De Loor, and P. Chevaillier. 2003. MASCARET: pedagogical multi-agents systems for virtual environment for training. In *Proceedings. 2003 International Conference on Cyberworlds*. IEEE Comput. Soc, Singapore, 423–430. https://doi.org/10.1109/CYBER.2003.1253485

[5] Deborah L. Butler and Philip H. Winne. 1995. Feedback and self-regulated learning: A theoretical synthesis. *Review of Educational Research* 65, 3 (1995), 245–281. https://doi.org/10.2307/1170684 Place: US Publisher: American Educational Research Assn.

[6] Myles Davaris, Sudanthi Wijewickrema, Yun Zhou, Patorn Piromchai, James Bailey, Gregor Kennedy, and Stephen O'Leary. 2019. The Importance of Automated Real-Time Performance Feedback in Virtual Reality Temporal Bone Surgery Training. In *Proceedings of the 2019 Artificial Intelligence in Education Conference*, Seiji Isotani, Eva Millán, Amy Ogan, Peter Hastings, Bruce McLaren, and Rose Luckin (Eds.). Springer International Publishing, Cham, 96–109.

[7] Daniele Di Mitri, Jan Schneider, and Hendrik Drachsler. 2022. Keep Me in the Loop: Real-Time Feedback with Multimodal Data. *Int J Artif Intell Educ* 32, 4 (Dec. 2022), 1093–1118. https://doi.org/10.1007/s40593-021-00281-z

[8] Alexandre Donzé and Oded Maler. 2010. Robust Satisfaction of Temporal Logic over Real-Valued Signals. In *FORMATS*. Lecture Notes in Computer Science, Vol. 6246. Springer, 92–106.

[9] Jon Echeverria and Olga C. Santos. 2021. KUMITRON: Artificial Intelligence System to Monitor Karate Fights that Synchronize Aerial Images with Physiological and Inertial Signals. In *Companion Proceedings of the 26th International Conference on Intelligent User Interfaces* (, College Station, TX, USA,) *(IUI '21 Companion)*. Association for Computing Machinery, New York, NY, USA, 37–39. https://doi.org/10.1145/3397482.3450730

[10] Jon Echeverria and Olga C. Santos. 2021. Toward Modeling Psychomotor Performance in Karate Combats Using Computer Vision Pose Estimation. *Sensors* 21, 24 (2021), 27 pages. https://doi.org/10.3390/s21248378

[11] Rita A. Gregory and Terrance Ward. 2000. Work Force Characteristics in a Robot Driven Construction Industry. In *Proceedings of the 17th IAARC/CIB/IEEE/IFAC/IFR International Symposium on Automation and Robotics in Construction*. International Association for Automation and Robotics in Construction (IAARC), Taipei, Taiwan, 1–5. https://doi.org/10.22260/ISARC2000/0072 ISSN: 2413-5844.

[12] Michael Haslgrübler, Benedikt Gollan, Christian Thomay, Alois Ferscha, and Josef Heftberger. 2019. Towards skill recognition using eye-hand coordination in industrial production. In *Proceedings of the 12th ACM International Conference on PErvasive Technologies Related to Assistive Environments* (Rhodes, Greece) *(PETRA '19)*. Association for Computing Machinery, New York, NY, USA, 11–20. https://doi.org/10.1145/3316782.3316784

[13] Thanos Hatziapostolou and Iraklis Paraskakis. 2010. Enhancing the Impact of Formative Feedback on Student Learning Through an Online Feedback System. *Electronic Journal of e-Learning* 8, 2 (2010), 111 – 122.

[14] Michael Henderson, Michael Phillips, Tracii Ryan, David Boud, Phillip Dawson, Elizabeth Molloy, and Paige Mahoney. 2019. Conditions that enable effective feedback. *Higher Education Research & Development* 38, 7 (Nov. 2019), 1401–1416. https://doi.org/10.1080/07294360.2019.1657807

[15] James Hutson and Jason Ceballos. 2023. Rethinking Education in the Age of AI: The Importance of Developing Durable Skills in the Industry 4.0. *Journal of Information Economics* 1, 2 (July 2023), 26–35. https://doi.org/10.58567/jie01020002

[16] Emily Jensen, Bradley Hayes, and Sriram Sankaranarayanan. 2023. More Than a Number: A Multi-dimensional Framework For Automatically Assessing Human Teleoperation Skill. In *Companion of the 2023 ACM/IEEE International Conference on Human-Robot Interaction*. ACM, Stockholm Sweden, 653–657. https://doi.org/10.1145/3568294.3580167

[17] Emily Jensen, Sriram Sankaranarayanan, and Bradley Hayes. 2024. Large Language Models Enable Automated Formative Feedback in Human-Robot Interaction Tasks. In *Human – Large Language Model Interaction Workshop*. ACM/IEEE, Boulder, CO, 2 pages.

[18] Tiina Korhonen, Timo Lindqvist, Joakim Laine, and Kai Hakkarainen. 2023. Training Hard Skills in Virtual Reality: Developing a Theoretical Framework for AI-Based Immersive Learning. In *AI in Learning: Designing the Future*, Hannele Niemi, Roy D. Pea, and Yu Lu (Eds.). Springer International Publishing, Cham, 195–213. https://doi.org/10.1007/978-3-031-09687-7_12

[19] Joakim Laine, Timo Lindqvist, Tiina Korhonen, and Kai Hakkarainen. 2022. Systematic Review of Intelligent Tutoring Systems for Hard Skills Training in Virtual Reality Environments. *International Journal of Technology in Education and Science* 6, 2 (May 2022), 178–203. https://doi.org/10.46328/ijtes.348 Number: 2.

[20] Ramona Diana Leon. 2023. Employees' reskilling and upskilling for industry 5.0: Selecting the best professional development programmes. *Technology in Society* 75 (Nov. 2023), 102393. https://doi.org/10.1016/j.techsoc.2023.102393

[21] Ling Li. 2022. Reskilling and Upskilling the Future-ready Workforce for Industry 4.0 and Beyond. *Inf Syst Front* 2022 (July 2022), 16 pages. https://doi.org/10.1007/s10796-022-10308-y

[22] Vanda Luengo and Dima Mufti-Alchawafa. 2013. Target the controls during the problem solving activity, a process to produce adapted epistemic feedbacks in ill- defined domains.. In *Formative Feedback in Interactive Learning Environments (FFILE) Conference*. Springer, Memphis, USA, 8.

[23] Cecilia Martinez, Ramiro Serra, Prem Sundaramoorthy, Thomas Booij, Cornelis Vertegaal, Zahra Bounik, Kevin Van Hastenberg, and Mark Bentum. 2023. Content-Focused Formative Feedback Combining Achievement, Qualitative and Learning Analytics Data. *Education Sciences* 13, 10 (Oct. 2023), 1014. https://doi.org/10.3390/educsci13101014

[24] Khaleel Asyraaf Mat Sanusi, Daniele Di Mitri, Bibeg Limbu, and Roland Klemke. 2021. Table Tennis Tutor: Forehand Strokes Classification Based on Multimodal Data and Neural Networks. *Sensors* 21, 9 (2021), 18 pages. https://doi.org/10.3390/s21093121

[25] Sofia Morandini, Federico Fraboni, Marco De Angelis, Gabriele Puzzo, Davide Diusino, and Luca Pietrantoni. 2023. The Impact of Artificial Intelligence on Workers' Skills: Upskilling and Reskilling in Organisations. *Informing Science: The International Journal of an Emerging Transdiscipline* 26 (Feb. 2023), 039–068. https://www.informingscience.org/Publications/5078

[26] Alexandra Moringen, Sören Rüttgers, Luisa Zintgraf, Jason Friedman, and Helge Ritter. 2021. Optimizing piano practice with a utility-based scaffold. arXiv:2106.12937 [cs.HC]

[27] Ana Moya, Leire Bastida, Pablo Aguirrezabal, Matteo Pantano, and Patricia Abril-Jiménez. 2023. Augmented Reality for Supporting Workers in Human–Robot Collaboration. *MTI* 7, 4 (April 2023), 40. https://doi.org/10.3390/mti7040040

[28] Abheek Anjan Mukherjee, Alok Raj, and Shikha Aggarwal. 2023. Identification of barriers and their mitigation strategies for industry 5.0 implementation in emerging economies. *International Journal of Production Economics* 257 (2023), 108770. https://doi.org/10.1016/j.ijpe.2023.108770

[29] Delwyn Nicholls, Linda Sweet, and Jon Hyett. 2014. Psychomotor Skills in Medical Ultrasound Imaging. *Journal of Ultrasound in Medicine* 33, 8 (2014), 1349–1352. https://doi.org/10.7863/ultra.33.8.1349 arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.7863/ultra.33.8.1349

[30] OpenAI. 2023. GPT-4V(ision) technical work and authors. https://openai.com/contributions/gpt-4v/

[31] Alberto Casas Ortiz. 2020. *Capturing, Modelling, Analyzing and providing Feedback in Martial Arts with Artificial Intelligence to support Psychomotor Learning Activities*. Master's thesis. Universidad Nacional de Educación a Distancia.

[32] Olga Pishchukhina and Angela Allen. 2021. Supporting learning in large classes: online formative assessment and automated feedback. In *2021 30th Annual Conference of the European Association for Education in Electrical and Information Engineering (EAEEIE)*. IEEE, Prague, Czech Republic, 1–4. https://doi.org/10.1109/EAEEIE50507.2021.9530953

[33] Miguel Portaz, Alberto Corbi, Alberto Casas-Ortiz, and Olga C. Santos. 2024. Exploring raw data transformations on inertial sensor data to model user expertise when learning psychomotor skills. *User Model User-Adap Inter* 2024 (April 2024), 43 pages. https://doi.org/10.1007/s11257-024-09393-2

[34] Daniele Pretolesi, Olivia Zechner, Daniel Garcia Guirao, Helmut Schrom-Feiertag, and Manfred Tscheligi. 2024. AI-Supported XR Training: Personalizing Medical First Responder Training. In *AI Technologies and Virtual Reality*, Kazumi Nakamatsu, Srikanta Patnaik, and Roumen Kountchev (Eds.). Springer Nature Singapore, Singapore, 343–356.

[35] Claudia Pérez-D'Arpino, Rebecca P. Khurshid, and Julie A. Shah. 2023. Experimental Assessment of Human-Robot Teaming for Multi-Step Remote Manipulation with Expert Operators. *J. Hum.-Robot Interact.* 2023 (Oct. 2023), 3618258. https://doi.org/10.1145/3618258

[36] Miguel Pérez-Ramírez, Norma J. Ontiveros-Hernández, Carlos A. Ochoa-Ortíz, José A. Hernández-Aguilar, and Benjamín E. Zayas-Pérez. 2016. Intelligent Tutoring Systems based on Virtual Reality for the Electrical Domain. *RCS* 122, 1 (Dec. 2016), 163–174. https://doi.org/10.13053/rcs-122-1-13

[37] Georg Rauter, Nicolas Gerig, Roland Sigrist, Robert Riener, and Peter Wolf. 2019. When a robot teaches humans: Automated feedback selection accelerates motor learning. *Sci. Robot.* 4, 27 (Feb. 2019), eaav1560. https://doi.org/10.1126/scirobotics.aav1560

[38] Jacob Rosen, Mika Sinanan, and Blake Hannaford. 2011. Objective Assessment of Surgical Skills. In *Surgical Robotics: Systems Applications and Visions*, Jacob Rosen, Blake Hannaford, and Richard M. Satava (Eds.). Springer US, Boston, MA, 619–649. https://doi.org/10.1007/978-1-4419-1126-1_25

[39] Olga C. Santos. 2016. Training the Body: The Potential of AIED to Support Personalized Motor Skills Learning. *Int J Artif Intell Educ* 26, 2 (June 2016), 730–755. https://doi.org/10.1007/s40593-016-0103-2

[40] Klaus Schwab and Saadia Zahidi. 2020. *The Future of Jobs Report 2020*. Technical Report. World Economic Forum. https://www3.weforum.org/docs/WEF_Future_of_Jobs_2020.pdf

[41] Florence H Sheehan, Shannon McConnaughey, Rosario Freeman, and R Eugene Zierler. 2019. Formative Assessment of Performance in Diagnostic Ultrasound Using Simulation and Quantitative and Objective Metrics. *Military Medicine* 184, Supplement_1 (March 2019), 386–391. https://doi.org/10.1093/milmed/usy388

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

[43] Aaron Steinfeld, Terrence Fong, David Kaber, Michael Lewis, Jean Scholtz, Alan Schultz, and Michael Goodrich. 2006. Common metrics for human-robot interaction. In *Proceedings of the 1st ACM SIGCHI/SIGART conference on Human-robot interaction*. ACM, Salt Lake City Utah USA, 33–40. https://doi.org/10.1145/1121241.1121249

[44] Ziheng Wang, Isabella Reed, and Ann Majewicz Fey. 2018. Toward Intuitive Teleoperation in Surgery: Human-Centric Evaluation of Teleoperation Algorithms for Robotic Needle Steering. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, Brisbane, QLD, 5799–5806. https://doi.org/10.1109/ICRA.2018.8460729

[45] David M. Williamson, Robert J. Mislevy, and Isaac I. Bejar (Eds.). 2006. *Automated Scoring of Complex Tasks in Computer-Based Testing*. Lawrence Erlbaum, Mahwah, New Jersey.

[46] Jing Yang, Juan Antonio Barragan, Jason Michael Farrow, Chandru P. Sundaram, Juan P. Wachs, and Denny Yu. 2024. An Adaptive Human-Robotic Interaction Architecture for Augmenting Surgery Performance Using Real-Time Workload Sensing—Demonstration of a Semi-autonomous Suction Tool. *Hum Factors* 66, 4 (April 2024), 1081–1102. https://doi.org/10.1177/00187208221129940 Publisher: SAGE Publications Inc.

[47] Madeleine S. Yuh, Ethan Rabb, Adam Thorpe, and Neera Jain. 2024. Using Reward Shaping to Train Cognitive–based Control Policies for Intelligent Tutoring Systems. In *2024 American Control Conference (ACC)*. IEEE, Toronto, ON, 8 pages.

[48] Madeleine Shuhn-Tsuan Yuh, Kendric Ray Ortiz, Kylie Sue Sommer-Kohrt, Meeko Oishi, and Neera Jain. 2024. Classification of Human Learning Stages via Kernel Distribution Embeddings. *IEEE Open Journal of Control Systems* 3 (2024), 102–117. https://doi.org/10.1109/OJCSYS.2023.3348704

[49] Vladimir Zotov and Eric Kramkowski. 2023. Moving-Target Intelligent Tutoring System for Marksmanship Training. *Int J Artif Intell Educ* 33, 4 (Dec. 2023), 817–842. https://doi.org/10.1007/s40593-022-00308-z