Int. J. Human–Computer Studies 183 (2024) 103180

Contents lists available at ScienceDirect

International Journal of Human - Computer Studies

journal homepage: www.elsevier.com/locate/ijhcs

Effects of adaptation accuracy and magnitude in affect-aware difficulty adaptation for the multi-attribute task battery

Vesna Dominika Novak [a,b,*], Dalton Hass [b], Mohammad Sohorab Hossain [a,b], Alexandria Fong Sowers [b], Joshua Dean Clapp [b]

[a] Department of Electrical and Computer Engineering, University of Cincinnati, 2851 Woodside Drive, Cincinnati, OH 45223, United States of America
[b] Department of Psychology, University of Wyoming, 1000 E University Avenue, Laramie WY 82071, United States of America

ARTICLE INFO

Keywords:
Affective computing
Difficulty adaptation
Human factors
Physiological computing
Psychophysiology
User experience

ABSTRACT

In affective computing, classification algorithms are used to recognize users' psychological states and adapt tasks to optimize user experience. However, classification is never perfect, and the relationship between adaptation accuracy and user experience remains understudied. It is also unclear whether the adaptation magnitude ('size' of action taken to influence user states) influences effects of adaptation accuracy. To evaluate impacts of adaptation accuracy (appropriate vs. inappropriate actions) and magnitude on user experience, we conducted a 'Wizard of Oz' study where 112 participants interacted with the Multi-Attribute Task Battery over three 11-minute intervals. An adaptation accuracy (50 % to 80 %) was preassigned for the first 11-minute interval, and accuracy increased by 10 % in each subsequent interval. Task difficulty changed every minute, and participant preferences for difficulty changes were assessed at the same time. Adaptation accuracy was artificially induced by fixing the percentage of times the difficulty changes matched participant preferences. Participants were also randomized to two magnitude conditions, with difficulty modified by 1 (low) or 3 (high) levels each minute. User experience metrics were assessed after each interval.

Analysis with latent growth models offered support for linear increases in user experience across increasing levels of adaptation accuracy. For each 10 % gain in accuracy, results indicate a 1.3 (95 % CI [.35, 2.20]) point increase in NASA Task Load Index scores (range 6–60), a 0.40 (95 % CI [.18, 0.57]) increase in effort/importance (range 2–14), and 0.48 (95 % CI [.24, 0.72]) increase in perceived competence (range 2–14). Furthermore, the effect of accuracy on Task Load Index scores was modulated by adaptation magnitude. No effects were observed for interest/enjoyment or pressure/tension. By providing quantitative estimates of effects of adaptation accuracy on user experience, the study provides guidelines for researchers and developers of affect-aware technologies. Furthermore, our methods could be adapted for use in other affective computing scenarios.

1. Introduction

1.1. Psychological state recognition and task adaptation in affective computing

In affective computing, machine learning algorithms are commonly used to recognize a user's psychological state (e.g., level of mental workload, engagement or frustration) based on measurements such as facial expressions or physiology (D'Mello et al., 2018; Picard et al., 2001). Usually, these algorithms are classifiers: they choose among multiple possible discrete states (Al-Nafjan et al., 2012; Samadiani et al., 2019; Shu et al., 2018). These states may be basic emotions such as fear, happiness or anger (Picard et al., 2001), or may be levels of a single variable: for example, low or high anxiety (S.M. Liu et al., 2009) or low or high enjoyment (Darzi and Novak, 2021). Most classifiers are supervised: they learn the relationship between inputs (measurements) and outputs (psychological states) based on previously collected training data from multiple human subjects. The training data are usually labelled using self-report questionnaires like the NASA task load index (TLX) (Hart and Staveland, 1988), which serve as the 'ground truth'.

Once the psychological state has been identified, actions can be taken to bring the user into a desirable state: for example, by adapting game difficulty to maximize player enjoyment or adapting the difficulty of learning materials to maximize student engagement (Aranha et al.,

[*] Corresponding author: Vesna Dominika Novak, University of Cincinnati, 2851 Woodside Drive, Cincinnati, OH 45223, United States of America.
E-mail address: novakdn@ucmail.uc.edu (V.D. Novak).

https://doi.org/10.1016/j.ijhcs.2023.103180
Received 16 March 2023; Received in revised form 15 August 2023; Accepted 1 November 2023
Available online 8 November 2023
1071-5819/© 2023 Elsevier Ltd. All rights reserved.

2021; Eldenfria and Al-Samarraie, 2019). Most affective computing systems classify the psychological state and have a single action defined for each state (e.g., "if state 1, then action 1″) (Aranha et al., 2021; D'Mello et al., 2018; D'Mello and Kory, 2015; Fairclough, 2017; Novak et al., 2012). For example, affect-aware games increase difficulty when anxiety is low and decrease it when anxiety is high, with each adaptation having a constant magnitude (increase/decrease by one level) (Darzi and Novak, 2021; C. Liu et al., 2009). Similarly, rehabilitation robots adapt the amount of assistance in response to patient workload, with each adaptation having a constant magnitude (increase/decrease by one level) (Koenig et al., 2011; Shirzad and Van der Loos, 2016; Xu et al., 2018). While more advanced adaptation rules have been presented (Cruz-Maya and Tapus, 2018; Liu et al., 2008), these still select among a few possible actions.

Despite its relative simplicity, affect-aware task adaptation (i.e., adaptation based on classified psychological state) has shown positive results. For example, affect-aware adaptation improves performance in unmanned vehicle control (Wilson and Russell, 2007), air traffic control (Aricò et al., 2016), and other high-stress tasks (Ung et al., 2018) compared to no feedback. Furthermore, it results in higher enjoyment in games (Liu et al., 2009), driving simulators (Bian et al., 2019) and rehabilitation (Xu et al., 2018) than adaptation based only on task performance. Finally, it results in higher enjoyment in games (Ewing et al., 2016; Nacke et al., 2011) and lower workload in resource management tasks (Bailey et al., 2006) than manual adaptation.

1.2. *The impact of classification and adaptation accuracy*

No affective computing system is perfect: it may incorrectly recognize the user's psychological state, consequently leading to inappropriate adaptation actions. Alternatively, it may recognize the correct psychological state but fail to take an action that would be appropriate for that state. For the type of affective computing system described in the previous section, we can thus define a classification accuracy (percentage of times correct psychological state recognized) and adaptation accuracy (percentage of times correct action taken).

Since most psychological state recognition classifiers rely on questionnaires as 'ground truth', a classifier is considered to have 100 % accuracy if it always outputs the psychological state self-reported by the user. In practice, classifiers never achieve 100 % accuracy: for example, our 2012 review of psychological state classification based on physiological measurements found accuracies mostly between 60 % and 90 % in two-class classification (e.g., low vs. high workload) and as low as 40 % in multiclass classification (e.g., low vs. medium vs. high workload) (Novak et al., 2012). Two 2019 reviews of classification using facial expressions similarly found multiclass classification accuracies as low as 50 % in real-world environments (Dewan et al., 2019; Samadiani et al., 2019). Furthermore, reviews of classification using electroencephalography (Al-Nafjan et al., 2017), general physiological signals (Shu et al., 2018), and posture and gestures (Stephens-Fripp et al., 2017) all found broad accuracy ranges, with very few studies reporting accuracies above 95 %.

An affective computing system can be considered to have 100 % adaptation accuracy if all its actions optimally guide users toward more appropriate psychological states. However, adaptation accuracy has been studied significantly less than classification accuracy. This may be because most state-of-the-art systems have a single adaptation action associated with each classified state (Aranha et al., 2021; D'Mello et al., 2018; D'Mello and Kory, 2015; Fairclough, 2017; Novak et al., 2012), making classification and adaptation closely linked. Nonetheless, high classification accuracy does not necessarily guarantee high adaptation accuracy. Since a single action is often predefined for each psychological state, that action may not always be optimal for that state – for example, an action defined for the state of "high workload" may be designed to increase automated assistance by a small amount, which may be appropriate for slightly excessive workload but not extremely excessive workload. Alternatively, since classification is often done over intervals of 2–5 min (Aranha et al., 2021; Novak et al., 2012), a system may not be able to recognize intense brief events (e.g., sudden extreme stress) until it is too late to act on them. Nonetheless, many affective computing studies either conflate classification and adaptation accuracy or study only classification accuracy.

It is generally assumed that higher classification and adaptation accuracies lead to better user experience. However, the actual relationship between accuracy and user experience in affect-aware adaptation is unclear, as also emphasized by other authors (Fairclough et al., 2015; Fairclough and Lotte, 2020). For example, our recent study found no improvement in user experience as a result of adding physiological measurements despite higher classification accuracy (Darzi et al., 2021). Without a clear idea of this relationship, it is difficult to make practical design decisions. For example, if adding another sensor increases classification accuracy by 5 %, what is the improvement in user experience and does it justify the added cost and complexity?

1.3. *Our previous wizard of Oz research*

If we assume that classification uses self-report data as 'ground truth' and that classification accuracy and adaptation accuracy are equal (as commonly done in affective computing), the relationship between accuracy and user experience in affect-aware adaptation could be studied using a 'Wizard of Oz' approach (Riek, 2012). Essentially, researchers could ask users how they would like to adapt the task (ground truth) and follow that preference a given percentage of the time to artificially create an adaptation accuracy. For example, if we ask a user about game difficulty 10 times and they always tell us to increase difficulty, we could artificially induce 80 % accuracy by increasing difficulty 8 times and decreasing it twice. This would allow the relationship between accuracy and user experience to be studied in a systematic manner, as different accuracies could be easily induced without complex signal processing and without random elements such as fluctuations in accuracy due to inter- or intrapersonal variability.

Co-author Novak previously conducted two studies using this Wizard of Oz approach. In the first study (Novak et al., 2014), participants played a game online, with each participant experiencing an adaptation accuracy between 50 % and 100 %. User satisfaction with the adaptation increased with adaptation accuracy as expected, but in-game fun surprisingly did not. However, the study had multiple methodological issues. For example, due to design flaws, both the interval between adaptation actions and the total number of adaptation actions were highly variable between participants. Additionally, since the study was done online, no sensors were involved, and it is unknown whether participants paid attention to study instructions. Finally, there was a dropout rate of over 40 %, with many more dropouts at lower adaptation accuracies.

To address these weaknesses, Novak carried out a second study (McCrea et al., 2017) where participants played a game on a lab computer while wearing an electroencephalography sensor. Each participant played for two 7-minute intervals, with difficulty adapted every 60 s. They experienced a different adaptation accuracy in each interval (among 33 %, 50 %, 66.7 %, 83.3 % and 100 %). Participants were told that difficulty was adapted based on sensor readings and that the two intervals corresponded to two different machine learning algorithms, but adaptation was actually done using the Wizard of Oz approach: participants could ask to decrease, increase or not change difficulty, and the game followed their preference a predefined percentage of the time. In this study, both satisfaction with adaptation as well as in-game fun were correlated with adaptation accuracy. Furthermore, while participants who experienced two very different accuracies (e.g., 50 % and 100 %) could reliably tell the difference between them, participants who experienced more similar accuracies (e.g., 83.3 % and 100 %) could not, suggesting that small accuracy differences were not overtly perceived by users.

1.4. Contribution of current study

While the above in-person study (McCrea et al., 2017) provided valuable insights into the relationship between adaptation accuracy and user experience, it still had limitations. For example, game difficulty was always adapted by one level, and it is unclear whether adapting by multiple levels would have a different effect. Additionally, it restricted within-person assessment to two randomly assigned accuracy conditions, limiting formal estimation of change in user experience following incremental change in adaptation accuracy. Furthermore, both above studies were performed using games, and it is unclear if results would be different in a more serious context. In the current study, a larger number of participants thus interacted with a computer-based multitasking scenario for three 11-minute intervals, with each successive interval administered at an artificially induced 10 % increase in adaptation accuracy. The initial accuracy and the magnitude of adaptation actions were randomized among participants, allowing us to obtain a more extensive look into the relationship between adaptation accuracy and user experience in affective computing.

2. Materials and methods

The study had two main goals:

1 Determine how adaptation accuracy influences user experience in an affective computing system involving a computer-based multitasking scenario. While higher adaptation accuracy is expected to result in better user experience, the relationship between the two is still unclear.
2 Determine how the magnitude of adaptation actions influences the change in user experience as a function of adaptation accuracy. High-magnitude adaptation actions might lead to faster convergence to a desirable user state, but only if adaptation accuracy is high enough. Thus, higher adaptation magnitude may lead to more aggressive change in user experience with increasing accuracy.

To achieve these goals, each participant performed a computer-based multitasking scenario for three 11-minute intervals. In the first of these three intervals, a Wizard of Oz approach was used to artificially induce an adaptation accuracy of 50 %, 60 %, 70 %, or 80 % (randomized across participants). The artificially induced adaptation accuracy then increased by 10 % with each successive interval. For example, participants who experienced 50 % accuracy in their first interval then experienced 60 % in their second interval and 70 % in their third interval; participants who first experienced 80 % then experienced 90 % and finally 100 %. This resulted in overlapping, within-group trials covering the full range of 50 % to 100 % accuracy levels. This approach, commonly known as a rolling panel design (Frees, 2004), permitted the estimation of a continuous growth function for user experience without necessitating that all participants experience all accuracies.

In addition to being randomized to one of four initial accuracies, participants were also randomized to one of two adaptation magnitude conditions. In the low-magnitude condition, any adaptation action changed difficulty by 1 level. In the high-magnitude condition, any adaptation action changed difficulty by 3 levels.

2.1. Participants

112 students (72 women, 40 men, no participants identified as nonbinary) were recruited from undergraduate psychology courses at the University of Wyoming and given course credit for study participation. Students were 20.3 ± 2.9 years old (mean ± standard deviation). When asked how often they play computer games (options: never, less than 2 h/week, 2–5 h/week, 6–10 h/week, 11–20 h/week, 20+ hours/week), 37.5 % reported never playing and another 44.5 % indicated 5 or fewer total hours of game play per week. When asked how difficult students preferred games to be on a 1 (not at all) to 7 (very difficult) scale, their preference was 3.6 ± 1.1. Participants also self-reported Big Five personality traits using the Ten Item Personality Inventory (Gosling et al., 2003); on a scale of 2–14, scores were 8.7 ± 2.7 for extraversion, 9.7 ± 2.1 for agreeableness, 10.7 ± 2.1 for conscientiousness, 9.2 ± 2.5 for emotional stability, and 10.9 ± 1.9 for openness to experiences. Only 3.6 % of the sample reported previously participating in affective computing research.

Approximately half the participants ($n = 54$) were randomized to the high-magnitude adaptation condition. A total of 30, 31, 19, and 32 participants initiated the scenario at 50 %, 60 %, 70 %, and 80 % accuracy, respectively.

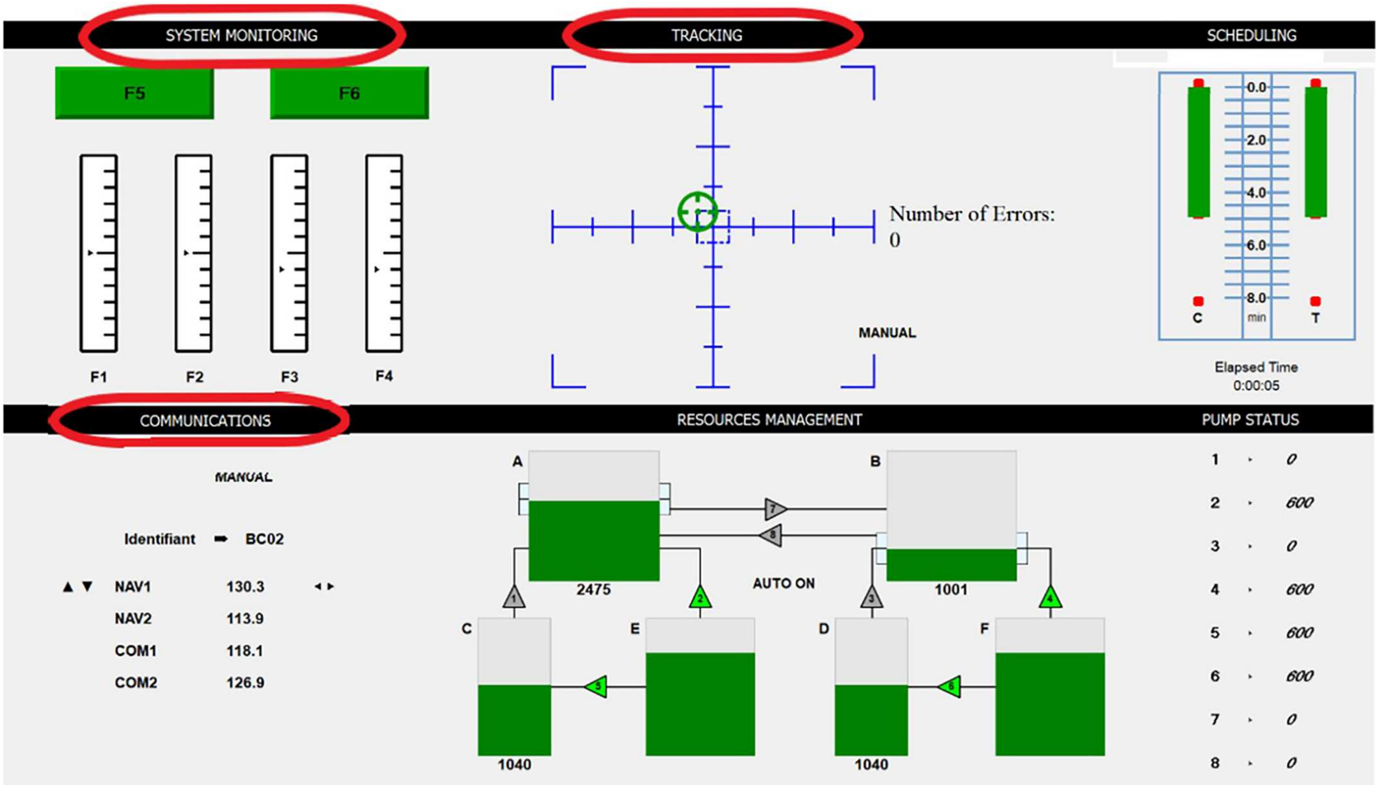
2.2. Scenario

The scenario used for the study was the OpenMATB (Cegarra et al., 2020), an open-source version of the NASA Multi-Attribute Task Battery (Santiago-Espada et al., 2011), a multitasking scenario commonly used to induce workload in affective computing. It was performed on a personal computer using a keyboard, joystick and headphones. A screenshot is shown in Fig. 1.

The standard OpenMATB includes six screen sections: system monitoring, tracking, scheduling, communications, resources management, and pump status. For our study, the scheduling, resources management and pump status sections ran automatically, and participants did not have to interact with them. Additionally, a “Number of errors” counter (not present in the standard OpenMATB) was added near the middle of the screen. The remaining sections were:

- Tracking (Fig. 1, upper middle): Participants must use the joystick to keep the green reticle inside the small central square. The reticle drifts out of the square if not actively maintained, with the speed and unpredictability of drift dependent on difficulty level. If the reticle stays outside the square for a few seconds, it flashes red and the error count increases by 1.
- System monitoring (Fig. 1, upper left): Four vertical columns include arrows that start near the center, but gradually move toward the top or bottom of the column. If an arrow gets too close to the top/bottom, the participant must hit the button corresponding to that column (F1-F4) to reset it to the center. The two green lights in the top left occasionally turn yellow; the participant must then press the corresponding button (F5, F6) to reset it. If a button is not pressed in time, the error count increases by 1 and the light/column flashes red then resets to the center. If a button is pressed unnecessarily, the error count also increases by 1 and that light/column flashes red. The difficulty level affects how often the buttons must be pressed.
- Communications (Fig. 1, lower left): Periodically, spoken instructions come over the headphones in the form of “[I*dentifiant*]”, turn your [channel] to [number].” If the identifiant part corresponds to the identifiant shown on the screen, the instruction should be followed; otherwise, it should be ignored. To follow the instruction, participants must use the up/down keyboard buttons to navigate to the correct channel (NAV1, NAV2, COM1, COM2) and then the left/right buttons to change the number to the requested one. If the correct number is not set in time, the error count increases by 1. The difficulty level affects how often instructions are spoken.

Ten difficulty levels were implemented for the OpenMATB, with level 1 being very easy (few monitoring/communications events, tracking reticle barely moves) and level 10 being very difficult (frequent events, reticle moves rapidly and unpredictably). To allow easier reproduction and expansion of our study, this version of the OpenMATB has been uploaded to Zenodo (Novak et al., 2022); it can be run using any Python interpreter and includes detailed settings for all ten difficulty levels. Table 1 shows individual difficulty settings in the ten difficulty levels.



MANUAL

Identifiant ➡ BC02

NAV1	130.3
NAV2	113.9
COM1	118.1
COM2	126.9

1	0
2	600
3	0
4	600
5	600
6	600
7	0
8	0

Fig. 1. Screenshot of the OpenMATB used in our study. While the original OpenMATB includes six display elements, only the three subtasks circled in red were assigned to participants in our study. The "Number of errors" counter near the middle was added to the OpenMATB for our study.

Table 1
OpenMATB difficulty settings for the three subtasks (tracking, monitoring, communications). In the tracking task, "drift amount" is a variable affecting the speed and unpredictability of drift and is defined in the original OpenMATB code; values can be compared to each other but have no absolute interpretation. In the monitoring task, the number of events represents how many times F1-F6 need to be pressed by the user. In the communications task, the number of events represents the number of instructions that come over the headphones; "for user" events must be attended to by the participant while "for others" events should be ignored.

Difficulty level	Tracking drift amount	Number of monitoring events	Number of communications events	
			for user	for others
1	0.07	2	1	0
2	0.08	3	0	2
3	0.09	4	1	1
4	0.10	5	1	1
5	0.11	6	1	1
6	0.12	7	2	1
7	0.13	8	2	1
8	0.14	9	3	0
9	0.15	12	3	0
10	0.16	15	3	0

2.3. *Study protocol*

A flowchart of the study protocol is shown in Fig. 2. Upon arrival, participants were told that the goal of the study was to test three different affect-aware systems that adapted the difficulty of a computer-based scenario based on physiological measurements. They were told that their own preferences about scenario difficulty would be collected but would not be used to adapt difficulty – only to verify system performance after the session. The study protocol and sensors were explained, and participants signed an informed consent form. They then filled out initial questionnaires (Section 2.1).

Participants sat at a computer, put on headphones, and self-applied sham physiological sensors: three disposable electrodes on the torso to record the electrocardiogram and reusable dry electrodes on the distal phalanges of the forefinger and middle finger of the nondominant hand to record skin conductance. No data were collected from these sensors. The scenario (Section 2.2) was then started, the individual scenario sections were explained, and participants completed a 5-minute practice interval with the scenario at difficulty level 5 of 10. After the practice, the experimenter answered any questions. To maintain the Wizard of Oz illusion, participants then sat quietly for 60 s to "obtain baseline physiological data". Finally, a camera-based eye tracker under the screen was "calibrated" by having participants look at each corner of the screen for a few seconds. No data were collected from the eye tracker either.

After calibration, participants interacted with the scenario for three 11-minute intervals. In each interval, the scenario started at difficulty level 5. Every 60 s, the scenario paused to ask participants how they would like to change difficulty (options: increase, decrease, don't change). Once participants input their choice, the difficulty was adapted according to the assigned accuracy and magnitude condition, and the scenario continued. The adaptation action actually taken by the scenario was not explicitly told to participants, but could be inferred from visible changes in difficulty.

After each 11-minute interval, participants filled out two questionnaires: the Intrinsic Motivation Inventory (IMI - same 8-item version as in our previous work (Goršič et al., 2017)) and NASA TLX (Hart and Staveland, 1988). After the final 11-minute interval ended, participants were asked "How much did you like the three difficulty adaptation algorithms" and rated each on a visual analog scale (VAS) from "did not like at all" to "liked very much". Participants could see all three VAS answers simultaneously and were encouraged to consider the three intervals relative to each other. While this is not a validated questionnaire,

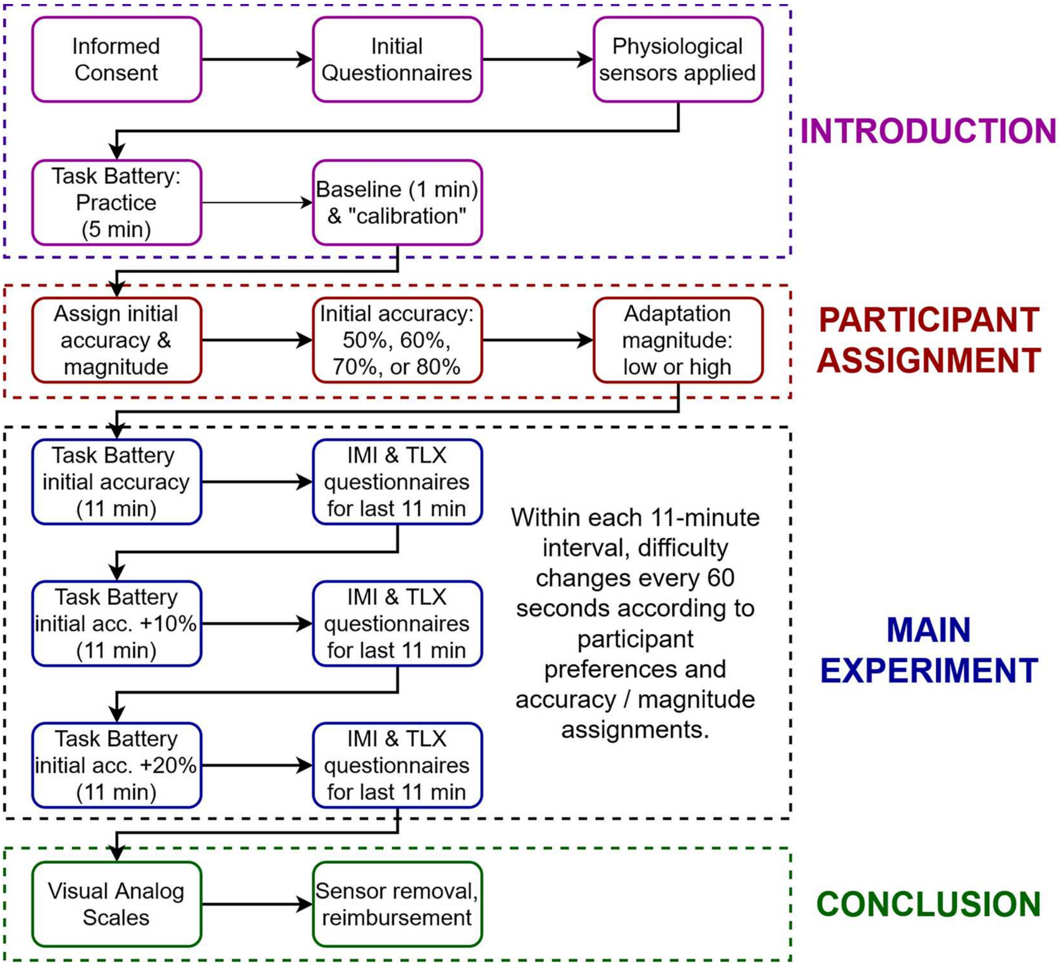

Informed Consent
Initial Questionnaires
Physiological sensors applied
INTRODUCTION
Task Battery: Practice (5 min)
Baseline (1 min) & "calibration"
Assign initial accuracy & magnitude
Initial accuracy: 50%, 60%, 70%, or 80%
Adaptation magnitude: low or high
PARTICIPANT ASSIGNMENT
Task Battery initial accuracy (11 min)
IMI & TLX questionnaires for last 11 min
Within each 11-minute interval, difficulty changes every 60 seconds according to participant preferences and accuracy / magnitude assignments.
Task Battery initial acc. +10% (11 min)
IMI & TLX questionnaires for last 11 min
MAIN EXPERIMENT
Task Battery initial acc. +20% (11 min)
IMI & TLX questionnaires for last 11 min
Visual Analog Scales
Sensor removal, reimbursement
CONCLUSION

Fig. 2. Study protocol flowchart. After putting on sensors and practicing the Task Battery (Introduction), participants are assigned to one of four possible accuracies and one of two possible magnitudes. They then interact with the Task Battery for three 11-minute intervals, with each interval at a 10 % higher accuracy. Intrinsic Motivation Inventory (IMI) and Task Load Index (TLX) questionnaires are filled out after each 11-minute interval, and final Visual Analog Scale questionnaires are filled out at the end.

a similar version was used in our prior work (McCrea et al., 2017). Finally, participants removed the sensors, were thanked for their participation, and received course credit.

2.4. Induction of accuracies

As mentioned, participants began the first 11-minute interval with initial accuracies ranging from 50 % to 80 % (with the initial accuracy preassigned to each participant), and the accuracy was increased by 10 % in each subsequent interval. Each participant was exposed to three accuracies since our previous work suggested that participants are poor at rating individual accuracies in isolation but are able to compare them to each other (McCrea et al., 2017). As the actual classification and adaptation accuracy of an affective computing system varies between participants as well as over time, we instead used the same Wizard of Oz approach from our previous work (McCrea et al., 2017; Novak et al., 2014) to artificially induce adaptation accuracies.

Participants were asked how they would like to change difficulty 10 times over the 11-minute interval. Since the participant's own preference is often considered to be "correct" in affective computing (Aranha et al., 2021; D. Novak et al., 2012), an accuracy of 100 % could be artificially induced by simply following the participant's preference all 10 times. On the other hand, an accuracy of 50 % could be induced by following the participant's preference 5 of 10 times and not following it the other 5 times. Thus, to induce a particular accuracy, the 10 adaptation actions in the 11-minute interval were predefined to agree or disagree with the participant's preference as follows:

- 50 % accuracy: System disagreed with participant after first, third, fourth, eighth and tenth minute.
- 60 %: System disagreed with participant after second, fifth, sixth, and ninth minute.
- 70 %: System disagreed with participant after second, fifth, and eighth minute.
- 80 %: System disagreed with participant after third and seventh minute.
- 90 %: System disagreed with participant after seventh minute.
- 100 %: System never disagreed.

These patterns were the same for all participants experiencing a given accuracy, ensuring that the exact desired accuracy was always induced.

When agreeing with the participant, the system adapted difficulty as requested by the participant (increase, decrease, don't change). When disagreeing, if the participant had asked to increase/decrease difficulty, the system instead decreased/increased it. If the participant had asked to keep difficulty the same, the system changed it in a random direction. For participants in the low-magnitude adaptation condition, any increase/decrease changed difficulty by 1 level; for participants in the high-magnitude condition, any increase/decrease changed it by 3 levels.

As the scenario had ten difficulty levels, the adaptation was capped so that if an increase would change difficulty above 10, it instead set it to 10; similarly, if a decrease would decrease it below 1, it would instead set it to 1. All this behavior is implemented and reusable in the version of OpenMATB shared to Zenodo (Novak et al., 2022).

As a result of this implementation, two different participants experiencing the same accuracy for 11 min would go through the same difficulty level sequence if they input the same sequence of desired actions and never requested to keep difficulty the same. A difficulty increase followed by a decrease (or vice versa) would return difficulty to the original level as long as neither action attempted to exceed the difficulty caps at levels 1 and 10.

2.5. Data analysis

There were five primary outcome variables summarizing self-reported participant experience with the scenario: the NASA TLX score, which represents total workload on a 6–60 scale (obtained as a raw sum of all six TLX items, with performance item reversed (Hart and Staveland, 1988)), and the four IMI scales (interest/enjoyment, effort/importance, perceived competence, and pressure/tension – each on a 2–14 scale (Goršič et al., 2017)). In addition, there was one secondary outcome variable describing scenario behavior rather than user experience: mean difficulty, defined as the mean difficulty level over all 11 min of an interval. Each variable was analyzed separately.

Continuous change in participant experience as a function of increasing adaptation accuracy was examined through a series of latent growth curve (LGC) analyses. These models assume that patterns of change in responding arise as the result of latent, unobserved growth parameters that drive observed scores (Bollen and Curran, 2006). For these analyses, individual trajectories of participant experience are initially combined to produce a baseline model of change for the sample as a whole. Variability in estimates of growth parameters set for the baseline model (intercept, slope) are then evaluated, with evidence of significant variability across participant-specific trajectories indicating the potential for moderators of initial status and rate of growth (i.e., individual difference factors that may impact patterns of change).

Analyses were conducted using a stepped approach (Bollen and Curran, 2006). Initial baseline models were used to estimate average trajectories of participant experience across increasing levels of adaptation accuracy. Variances for intercept (i.e., starting value at 50 % accuracy) and slope (i.e., change in experience as a function of adaptation accuracy) parameters in baseline models for each outcome were examined to assess for potential differences across respondents, suggesting participant-specific trajectories that may be attributable to condition-level differences in the magnitude of adaptation actions (low vs. high). For models with significant variance estimates, magnitude condition was included as a potential moderator of corresponding intercept and slope parameters (i.e., exploring whether initial response and/or rate of change across accuracy values differed across high- and low-magnitude adaptation conditions).

Analyses were conducted in MPlus 8.8 (Muthén and Muthén, 2017) using full information maximum likelihood estimation. Full information estimators are capable of accommodating cases with partially missing values, permitting the generation of growth parameters in rolling panel designs such as the one used for this study. Parameters are estimated using all information available from the set and remain unbiased when missingness is unrelated to other variables in the model (Arbuckle et al., 1996). Because missing values in rolling panel designs are the result of experimental randomization as opposed to person-level factors (e.g., dropout due to frustration with low-accuracy systems), data remain amenable to full information maximum likelihood estimation.

For the “How much did you like the three difficulty adaptation algorithms” question at the end of the session, VAS scores were analyzed using a 2 (magnitude: low, high) x 4 (initial accuracy: 50, 60, 70, 80) x 3 (trial: 1, 2, 3) mixed-factors analysis of variance (ANOVA) with repeated measures on the final factor. Interaction effects in this model were used to assess whether the pattern of average liking across trials was dependent upon initial accuracy and/or adaptation condition.

3. Results

Descriptive statistics for mean difficulty and the number of times difficulty was increased, decreased or left unchanged in each 11-minute interval are shown in Table 2 for all adaptation accuracies and both magnitude conditions. Note that increase / decrease / don’t change refer to the actual difficulty changes, not requests made by participants (which the system could disagree with). Descriptive statistics for NASA TLX and IMI scores across adaptation accuracies are then presented in Table 3.

3.1. Mean difficulty

Linear growth parameters (intercept = 5.11, $p < .001$; slope = 0.30, $p < .001$) estimated a 0.30 unit increase in average task difficulty for every 10 % increase in accuracy from 50 % to 100 % trials. A 95 % confidence bound suggest population estimates for growth ranging from a 0.17 to a 0.42 unit increase per 10 % improvement in adaptation accuracy ($CI_{95\ \%}$ [.17, 0.42]). Results did not provide support for variability across participants with respect to average difficulty at 50 % accuracy ($p = .278$) or in change across trials ($p = .637$) suggesting an overall stable trajectory across respondents and magnitude condition.

3.2. NASA task load index

Growth parameters for the baseline model (intercept = 32.80, $p < .001$; slope = 1.01, $p = .001$) identified an aggregate trajectory of increasing TLX scores in response to increasing adaptation accuracy. Results indicated significant variability across respondents at initial (50 %) adaptation accuracy ($p = .037$), suggesting the potential for differences in TLX scores as a result of magnitude condition. Variability for slope, however, did not achieve statistical significance ($p = .330$), suggesting similar trajectories of increasing scores across users.

Table 2
Means ± standard deviations for mean difficulty and the number of times that difficulty was increased, decreased or left unchanged in each 11-minute interval, separately for low- and high-magnitude conditions.

Accuracy	Magnitude	n	Mean difficulty	Number of adaptation actions		
				increase	don’t change	decrease
50 %	low	15	5.3 ± 0.9	3.3 ± 1.4	3.7 ± 2.1	3.0 ± 1.3
	high	15	5.9 ± 1.5	2.5 ± 1.2	4.6 ± 2.5	2.9 ± 1.4
60 %	low	30	5.5 ± 1.3	3.7 ± 1.3	3.6 ± 1.6	2.7 ± 1.3
	high	31	5.6 ± 1.3	2.5 ± 1.2	4.8 ± 2.0	2.7 ± 1.0
70 %	low	41	5.4 ± 1.4	3.3 ± 1.6	4.1 ± 1.8	2.6 ± 1.1
	high	39	5.6 ± 1.7	2.3 ± 1.0	5.2 ± 1.7	2.5 ± 1.0
80 %	low	43	5.5 ± 1.3	3.1 ± 1.8	4.7 ± 1.9	2.2 ± 1.1
	high	39	5.7 ± 1.6	2.1 ± 0.8	5.6 ± 1.5	2.3 ± 1.0
90 %	low	28	6.0 ± 1.4	3.5 ± 1.9	4.6 ± 2.2	1.9 ± 1.2
	high	23	6.1 ± 1.6	2.0 ± 1.0	6.1 ± 1.9	1.9 ± 1.1
100 %	low	17	6.3 ± 1.6	3.3 ± 1.5	5.3 ± 2.2	1.5 ± 1.5
	high	15	6.6 ± 1.5	1.8 ± 1.3	6.6 ± 2.6	1.6 ± 1.4

Table 3
Means ± standard deviations for NASA Task Load Index (TLX) scores and the four scales of the Intrinsic Motivation Inventory at different adaptation accuracies. All scores are averaged across both low- and high-magnitude conditions.

Accuracy	n	NASA TLX	Interest/ Enjoyment	Effort/ Importance	Competence	Pressure/ Tension
50 %	30	32.2 ± 7.3	10.0 ± 2.5	11.4 ± 2.3	8.8 ± 2.6	9.6 ± 3.0
60 %	61	34.5 ± 7.0	10.0 ± 2.7	11.8 ± 2.0	8.4 ± 2.5	9.6 ± 2.9
70 %	80	34.8 ± 6.9	10.1 ± 2.8	11.9 ± 1.9	8.9 ± 2.5	9.7 ± 2.3
80 %	82	35.6 ± 7.6	10.3 ± 2.9	12.3 ± 1.8	9.1 ± 2.8	10.1 ± 2.8
90 %	51	37.3 ± 8.5	10.4 ± 2.9	12.6 ± 1.8	9.3 ± 2.8	10.6 ± 2.6
100 %	32	36.6 ± 9.5	10.3 ± 3.3	12.5 ± 2.1	9.6 ± 3.1	10.9 ± 2.1

Evaluation of the full model including magnitude condition as a predictor of initial TLX scores again provided support for growth in TLX score in response to increasing adaptation accuracy (intercept = 30.89, $p < .001$; slope = 1.28, $p = .005$). Based on these data, investigators should expect a 1.3 unit increase in TLX scores for every 10 % increase in adaptation accuracy in the OpenMATB scenario. Data-supported values for TLX growth range from approximately one half to slightly more than a 2-point TLX increase per 10 % change in adaptation accuracy ($CI_{95\ \%}$ [.35, 2.20]). Results also indicated support for an effect of magnitude condition at the initial 50 % accuracy ($b = 3.89$, $p = .022$) such that participants randomized to the high-magnitude condition returned higher TLX scores than those randomized to the low-magnitude condition (Table 4). As expected from variance effects in the baseline model, analyses did not indicate significant differences in growth as a function of magnitude condition ($p = .358$).

3.3. *Intrinsic motivation inventory*

The baseline for IMI interest/enjoyment scores did not provide evidence for growth as a function of increasing adaptation accuracy in these data (intercept = 10.14, $p < .001$; slope = 0.02, $p = .773$). Analyses did provide support for variability in corresponding intercept ($p < .001$) and slope ($p = .035$) parameters; however, evaluation of the full model (intercept = 9.92, $p < .001$; slope = 0.04, $p = .714$) did not indicate group-level differences in initial interest/enjoyment values or change in scores as a function of magnitude condition (all $p \geq .420$).

Evaluation of IMI effort did indicate linear growth in scores across increasing accuracy levels for the sample as a whole (intercept = 11.36, $p < .001$; slope = 0.30, $p < .001$). Variance estimates provided support for participant-level differences in intercept ($p < .001$) and slope ($p = .021$) although inclusion of magnitude condition in the full model did not ultimately indicate effects on either parameter (all $p \geq .052$). Estimates in the final model (intercept = 10.88, $p <0.001$; slope = 0.36, $p < .001$) suggest an expected 0.40 unit increase in IMI effort/importance for every 10 % increase in adaptation accuracy for individuals completing the OpenMATB scenario. Plausible values for expected growth range from a lower bound of 0.18 units per 10 % increase in accuracy to an upper limit of 0.57 ($CI_{95\ \%}$ [.18, 0.57]).

Baseline analysis for IMI competence identified a trajectory of increasing scores (intercept = 7.92, $p < 0.001$; slope = 0.44, $p < .001$) with evidence for individual differences in initial competency ratings ($p < .001$) but not for change across increasing accuracy ($p = .267$). Evaluation of the full model including magnitude condition continued to support an aggregate trajectory of increased confidence with increasing adaptation accuracy (intercept = 7.92, $p < .001$; slope = 0.48, $p < .001$). Data-supported estimates suggest an expected 0.48 unit increase in IMI competence for every 10 % increase in adaptation accuracy for individuals completing the OpenMATB scenario. Plausible values for expected growth range from a lower bound of 0.24 per 10 % increase in accuracy to an upper limit of 0.72 ($CI_{95\ \%}$ [.24, 0.72]).

Table 4
Means ± standard deviations for NASA Task Load Index (TLX) scores, separated into participants randomized to the low-magnitude condition and those randomized to the high-magnitude condition.

Accuracy	Low magnitude		High magnitude	
	n	NASA TLX	n	NASA TLX
50 %	15	30.3 ± 8.8	15	34.1 ± 5.1
60 %	30	32.8 ± 7.0	31	36.1 ± 6.7
70 %	41	33.3 ± 6.8	39	36.3 ± 6.7
80 %	43	34.4 ± 7.8	39	37.0 ± 7.1
90 %	28	37.1 ± 8.1	23	37.6 ± 9.1
100 %	17	37.2 ± 11.3	15	35.9 ± 7.1

Finally, the initial model for IMI pressure/tension did not detect evidence of growth as a function of increasing adaptation accuracy (intercept = 9.54, $p < .001$; slope = 0.18, $p = .133$). Variance estimates offered support for differences in initial starting values ($p < .001$) although slope estimates were relatively consistent across person-specific trajectories ($p = .322$). Evaluation of the final model with magnitude condition as a predictor also did not provide support for change as a function of adaptation accuracy (intercept = 9.21, $p <0.001$; slope = 0.27, $p = .051$). Magnitude condition was unrelated to intercept ($p = .305$) or slope ($p = .347$) parameters.

3.4. *Ratings at end of session*

Across all participants, VAS answers to the "How much did you like the three difficulty adaptation algorithms" question, converted to 0–100 values, were 67.3 ± 22.7 for the first algorithm experienced by participants, 70.5 ± 21.5 for the second, and 77.2 ± 18.3 for the third. The mixed-factors ANOVA provided strong support for differences in liking scores across successive algorithm trials ($p < .001$; $\eta_p^2 = 0.091$). Follow-up *t*-tests indicated increased liking for the final algorithm relative to both the first ($p < .001$, $d = 0.48$) and second ($p = .004$, $d = 0.35$). Differences in liking ratings across the first and second algorithms were not significantly different ($p = .183$, $d = 0.15$). Initial accuracy and magnitude conditions did not influence ratings (all $p \geq .160$).

4. Discussion

LGC analysis of mean difficulty indicates that higher accuracies lead to higher task difficulty, as seen in Table 2. This makes sense within the current OpenMATB scenario: the initial difficulty level (5) is not very difficult given some practice, so participants tended to want to increase difficulty over time, and a more accurate adaptation algorithm allowed them to reach the desired higher difficulties. As the primary goal of the Wizard-of-Oz affective computing system in this scenario was to adapt task difficulty, any effects of adaptation accuracy on user experience likely occur via changes in task difficulty; effects of accuracy cannot be separated from effects of difficulty.

4.1. *Positive results: TLX scores, effort/importance, perceived competence*

Results indicate that, as adaptation accuracy increases, NASA TLX and effort/importance scores increase as well. This makes sense given that higher accuracies result in higher task difficulty, which is expected to require higher task load and effort. Similar results were observed in our previous studies outside affective computing, where automated upward-trending difficulty adaptation resulted in higher effort (Goršič et al., 2017). Nonetheless, the current study goes a step further by estimating the increase in effort and TLX score that can be expected from an increase in adaptation accuracy.

The current study also demonstrates that a higher adaptation magnitude results in higher TLX scores even at low accuracies. We had previously hypothesized that higher adaptation magnitudes may allow even inaccurate adaptation algorithms to reach a desirable level faster (McCrea et al., 2017), and that does appear to be the case here. As seen in Table 2, participants randomized to the high-magnitude condition tended to experience fewer difficulty increases despite experiencing similar mean difficulty levels (e.g., 100 % accuracy: mean 3.3 increases for low and 1.8 increases for high magnitude), confirming this interpretation. However, this result is not necessarily generalizable: in other scenarios, a high adaptation magnitude may cause the adaptation algorithm to "overshoot" and bring the participant to an excessively high difficulty, which may have negative consequences.

Perceived competence also increases with accuracy, which may be simply due to time spent with the scenario (higher accuracies tended to appear later) but may be due to participants feeling like they are performing well at an appropriate difficulty as opposed to performing well at a low difficulty or failing at an overly high difficulty. In our previous work involving the IMI and difficulty adaptation outside affective computing, we did not find significant differences in competence as a result of adaptation (Goršič et al., 2017), but that study was conducted with a computer game rather than a "work-like" task.

4.2. *Negative results: interest/enjoyment, ratings at end of session*

Interest/enjoyment on the IMI did not increase with adaptation accuracy. The IMI is a popular questionnaire and found interest/enjoyment differences between adaptation algorithms in our previous study (Goršič et al., 2017), so positive effects were expected a priori. At the same time, one of our previous studies also found no difference in interest/enjoyment between intelligent adaptation and random adaptation within an affect-aware computer game (Darzi et al., 2021), so this is not an isolated result. One possible explanation for the negative interest/enjoyment result is that the OpenMATB scenario simply is not particularly fun regardless of difficulty, and performing it at a more appropriate difficulty thus does not make it much more fun. Even in affect-aware games, two adaptation accuracies that differ by 10–20 % result only in small differences in self-reported fun (McCrea et al., 2017), so it would make sense that the differences in a more "work-like" scenario would be even smaller. However, this should not be considered a weakness of the OpenMATB – while a more fun scenario may exhibit larger effects of adaptation accuracy on interest/enjoyment, it may exhibit smaller effects on other outcome variables.

Finally, ratings at the end of the session do indicate that participants prefer the later intervals and consequently higher accuracies over the lower accuracies. However, this may be simply due to increased familiarity with the scenario over time. Furthermore, there were no significant differences between different initial accuracies or magnitude conditions. This indicates that in the absence of prior experience with affect-aware adaptation, participants did not necessarily like an accurate algorithm more than an inaccurate one. This is similar to our previous study with affect-aware games, which found that participants are mostly able to recognize the more accurate of two different adaptation algorithms but are poor at evaluating adaptation accuracies in isolation (McCrea et al., 2017).

4.3. *Follow-up study: same accuracy in all three intervals*

After concluding the primary study, we were concerned that some positive results may have occurred simply because higher accuracies always appear later in the session when participants are already familiar with the scenario. As a small follow-up study that is not reported in detail, we thus recruited another 49 participants and randomized them to the same initial accuracy and magnitude conditions, but with the same adaptation accuracy in all three intervals (e.g., participants with an initial accuracy of 80 % experienced 80 % accuracy in the first, second and third interval). While the sample size was small, there were much smaller differences between the three intervals in TLX, IMI scores, and ratings at the end of the session. For example, while mean ratings of the three algorithms at the end of the session in the main study were 67.3 (first interval), 70.5 (second) and 77.2 (third), these ratings in the follow-up study were 70.5, 69.6, and 72.9. Thus, order effects are unlikely to account for most of the positive results observed in the current study.

4.4. *Practical implications*

Within a multitasking scenario, our study demonstrates that a more accurate adaptation algorithm is able to reach higher effort / task load levels more quickly and provides participants with higher perceived competence. Specifically, a 10 % increase in adaptation accuracy results in an expected 1.3-point increase on the TLX (range: 6–60), 0.40-point increase on IMI effort/importance (range: 2–14), and 0.48-point increase on IMI perceived competence (range 2–14). Table 4 also shows that large adaptation accuracies result in higher TLX scores, and Table 2 indicates that high-magnitude adaptation is able to reach higher difficulty levels more quickly even if it is inaccurate.

Though researchers have spent decades studying ways to improve classification and adaptation accuracy in affective computing, there is still little information about how improving these accuracies influences the user experience. While our results are scenario-specific to a large degree, they provide a "quick" estimate of the effects of accuracy improvements, allowing researchers and developers to more easily decide whether it is worth investing time and money into accuracy improvements. For example, if a developer estimates in advance that adding another sensor would improve accuracy by 5 %, they could use the results of our study to estimate that this sensor would likely increase participants' TLX scores by about 0.6 points and their IMI effort/importance by about 0.2 points. Furthermore, the TLX increase would likely not be less than 0.2 points and may be as large as 1 point as indicated by confidence intervals. While a value judgment would still need to be made on these TLX and IMI scores, it is likely easier than judging adaptation accuracies. Thus, our LGC analyses go beyond our previous Wizard of Oz work, which showed that increasing accuracy does improve user experience but did not estimate the degree of user experience improvement (McCrea et al., 2017).

Furthermore, though our results are context-specific, the study also provides a generalizable methodology that could be used to both study other variables in the same scenario as well as study other scenarios entirely. For example, if a developer wishes to evaluate the effects of adaptation accuracy in their own scenario, they could simply replace the OpenMATB with their own scenario and use the same protocol to identify expected accuracy effects. If a large sample of potential users is available, such a Wizard of Oz study would likely require much less effort than developing and testing affect-aware technologies for the scenario of interest. Other factors of interest could also be studied by expanding the methodology. For example, developers are often curious whether a more convenient sensor with a shorter setup time could still achieve a positive user experience even if it is less accurate, and this could be evaluated in our methodology by randomizing participants to either a "convenient sensor with short setup time" or a "obtrusive sensor with long setup time" condition and then including this condition in LGC analyses. We thus hope that our research will serve as a broad foundation for further Wizard of Oz evaluations in affective computing.

Finally, the lack of significant effects of initial adaptation accuracy in the ratings at the end of the session (Section 3.4) emphasizes a potentially important point: participants without prior affective computing experience may not be able to tell if an algorithm is "good" or "bad". Thus, exposing participants to only a single adaptation algorithm may not effectively identify differences in user experience due to different adaptation behavior. This was also observed in our prior Wizard of Oz work (McCrea et al., 2017), and may partially explain results of, e.g., our prior affective game study which found no differences in user experience between adaptation algorithms despite different adaptation accuracies (Darzi et al., 2021). This emphasizes the need for comparative studies in affective computing: prototype affect-aware technologies should not be evaluated in isolation, but should have their performance compared to that of a "benchmark" system within the same users.

4.5. Study limitations

Both this study and our previous work (McCrea et al., 2017; Novak et al., 2014) were done with scenarios where adaptation errors do not have serious consequences. While an erroneous adaptation action may briefly frustrate the user, this can always be reversed later. However, there are many real-world affect-aware systems where this may not be the case: for example, air traffic control systems where overloading the user may lead to fatal accidents (Aricò et al., 2016) or driver monitoring systems (Darzi et al., 2018; Sanghavi et al., 2023) where the potential consequence of a false negative (not detecting distraction/drowsiness, leading to an accident) may be much more severe than the potential consequence of a false positive (unnecessarily warning the user, resulting in annoyance). In these cases, we believe that changes in adaptation accuracy would have larger effects on user experience than observed in the current study. Furthermore, while the current study assumes linear effects of accuracy on user experience, a high-stakes scenario might have a more complex accuracy-experience relationship – for example, a "step-like" relationship where all accuracies below a high threshold (e.g., 95 %) are perceived as equally insufficient and accuracies above it are perceived as approximately equally good. Modifications could still be made to our study protocol to better approximate "critical" scenarios: for example, adding conditions where an adaptation error can immediately end the session and cause the participant to forfeit financial rewards. Still, since truly severe consequences cannot be implemented in Wizard of Oz studies, it is unclear whether our protocol could be used to evaluate affect-aware technologies for such applications.

Additionally, our Wizard of Oz research has treated participants as entirely passive: they perceive the affective adaptation but do not attempt to influence it in any way. While this is appropriate for many technologies, more transparent technologies could indicate why a particular adaptation action was taken: for example, a message saying "task difficulty was increased because your respiration rate is low". In such situations, participants may adapt their behavior to influence the technology's decision-making: for example, by breathing faster. Similar human behavior adaptation in response to predictable classification errors has been observed in other areas of human-machine interaction (S.M. Chase et al., 2009; Hargrove et al., 2010), but would need to be studied using a different paradigm. Related to this limitation, our results also cannot be transferred to affective computing technologies where the computer simply provides insight into the affective state and the user must decide what to do with that information. An example of this would be emotion regulation technologies (Sadka and Antle, 2022), where the emotion information is visualized to the user but the user must take action themselves. Adaptation magnitude would not be relevant in such a scenario, and classification accuracy could not be easily studied with a Wizard of Oz paradigm.

Finally, multiple aspects of the study design were fixed by experimenters, constraining the number of situations experienced by participants. First, task difficulty in our scenario always started at the same level (5), leading to an upward trend in difficulty over time and possible impacts on user experience (Nagle et al., 2016). Second, within a given accuracy setting, system errors were set at fixed points for all participants (Section 2.4). However, in applied settings, adaptation errors occurring at different points (early in performance vs. late) could have different effects on user performance. Finally, participants were always exposed to accuracies in increasing order (e.g., 70–80–90 %), which may introduce order effects. We considered randomizing these aspects of the design early in study development but chose not to do so given the lack of existing work in this area and the complexity of disentangling person-level effects from generalizable, group-level trajectories. However, as a result, it is not possible to study factors such as error timing and habituation. If future studies wish to examine these aspects, they could modify the protocol by, for example, starting each 11-minute interval at a random difficulty, having the system disagree with the participant at random times within each 11-minute interval, and exposing participants to accuracies in random orders (or at least in both increasing and decreasing orders).

5. Conclusion

Our study contributes to the growing body of research on the relationship between classification/adaptation accuracy and user experience in affective computing. Within a well-known multitasking scenario (the Multi-Attribute Task Battery), LGC analyses showed that increasing adaptation accuracy by 10 % is expected to increase self-reported NASA TLX scores by 1.3 points ($CI_{95\ \%}$ [.35, 2.20]) on a 6–60 scale, self-reported effort/importance by 0.40 points ($CI_{95\ \%}$ [.18, 0.57]) on a 2–14 scale, and self-reported competence by 0.48 points ($CI_{95\ \%}$ [.24, 0.72]) on a 2–14 scale. Furthermore, the effect of adaptation accuracy on TLX scores was modulated by adaptation magnitude. Adaptation accuracy had no effect on interest/enjoyment or pressure/tension, contrasting with previously studied game-like scenarios where adaptation accuracy increased enjoyment. Finally, participants' ratings of three different adaptation accuracies indicated that participants do perceive differences between these accuracies, but that inaccurate and accurate adaptation are likely to result in similar ratings if the participant has no prior experience with adaptation.

By providing quantitative estimates of the effect of adaptation accuracy and adaptation magnitude on user experience, the study provides guidelines for researchers and developers of affect-aware technologies. While the above estimates are specific to our own scenario and participants, the same Wizard of Oz methodology could be adapted for use with many scenarios in which adaptation errors do not have irreversible consequences. By conducting similar Wizard of Oz studies, developers could, for example, estimate whether improving adaptation accuracy would improve user experience in their own scenario and whether such improvements would offset the time and money needed to improve accuracy (e.g., by buying additional sensors). In the long term, we hope that Wizard of Oz evaluations in affective computing can serve as an early validation of these technologies prior to real-world evaluations of finished products, helping guide eventual broad adoption of the technologies.

CRediT authorship contribution statement

Vesna Dominika Novak: Conceptualization, Data curation, Funding acquisition, Methodology, Project administration, Resources, Supervision, Visualization, Writing – original draft, Writing – review & editing. **Dalton Hass:** Data curation, Investigation, Methodology, Software, Validation, Writing – review & editing. **Mohammad Sohorab Hossain:** Data curation, Investigation, Software, Writing – review & editing. **Alexandria Fong Sowers:** Formal analysis, Investigation, Validation, Writing – review & editing. **Joshua Dean Clapp:** Conceptualization, Data curation, Formal analysis, Funding acquisition, Methodology, Project administration, Resources, Supervision, Validation, Visualization, Writing – original draft, Writing – review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Code shared on Zenodo, with link to repository cited in manuscript (https://zenodo.org/records/7339589).

Acknowledgments

This work was supported by the National Science Foundation [grant numbers 2007908 and 2151464]. The sponsor had no involvement in any aspect of the study. The authors thank Maja Goršič, Carrie Faaberg, Johnathan Walker, and Cody Akin for their assistance with data collection.

References

Al-Nafjan, A., Hosny, M., Al-Ohali, Y., Al-Wabil, A., 2017. Review and classification of emotion recognition based on EEG brain-computer interface system research: a systematic review. Appl. Sci. (Switzerland) 7, 1239. https://doi.org/10.3390/app7121239.

Aranha, R.V., Correa, C.G., Nunes, F.L.S., 2021. Adapting software with affective computing: a systematic review. IEEE Trans Affect Comput 12 (4), 883–899. https://doi.org/10.1109/TAFFC.2019.2902379.

Arbuckle, J.L, Marcoulides, G.A., Schumacker, R.E, 1996. Full information estimation in the presence of incomplete data. Advanced Structural Equation Modeling: Issues and Techniques. Lawrence Erlbaum Associates, pp. 243–277.

Aricò, P., Borghini, G., Di Flumeri, G., Colosimo, A., Bonelli, S., Golfetti, A., Pozzi, S., Imbert, J.-P., Granger, G., Benhacene, R., Babiloni, F., 2016. Adaptive automation triggered by EEG-based mental workload index: a passive brain-computer interface application in realistic air traffic control environment. Front. Hum. Neurosci 10, 539. https://doi.org/10.3389/fnhum.2016.00539.

Bailey, N.R., Scerbo, M.W., Freeman, F.G., Mikulka, P.J., Scott, L.A., 2006. Comparison of a brain-based adaptive system and a manual adaptable system for invoking automation. Hum. Factors 48 (4), 693–709. https://doi.org/10.1518/001872006779166280.

Bian, D., Wade, J., Swanson, A., Weitlauf, A., Warren, Z., Sarkar, N., 2019. Design of a physiology-based adaptive virtual reality driving platform for individuals with ASD. ACM Trans. Access Comput 12 (1), 2.

Bollen, K.A., Curran, P.J., 2006. Latent Curve Models: A Structural Equation Perspective. John Wiley & Sons.

Cegarra, J., Valéry, B., Avril, E., Calmettes, C., Navarro, J., 2020. OpenMATB: a multi-attribute task battery promoting task customization, software extensibility and experiment replicability. Behav. Res. Methods 52, 1980–1990.

Chase, S.M., Schwartz, A.B., Kass, R.E., 2009. Bias, optimal linear estimation, and the differences between open-loop simulation and closed-loop performance of spiking-based brain-computer interface algorithms. Neural. Netw. 22, 1203–1213.

Cruz-Maya, A., Tapus, A., 2018. Adapting robot behavior using regulatory focus theory, user physiological state and task-performance information. In: RO-MAN 2018 - 27th IEEE International Symposium on Robot and Human Interactive Communication. https://doi.org/10.1109/ROMAN.2018.8525648.

D'Mello, S.K., Kory, J, 2015. A review and meta-analysis of multimodal affect detection systems. ACM Comput. Surv 47 (3), 43. https://doi.org/10.1145/2682899.

D'Mello, S., Kappas, A., Gratch, J, 2018. The affective computing approach to affect measurement. Emotion Rev. 10 (2), 174–183. https://doi.org/10.1177/1754073917696583.

Darzi, A., Gaweesh, S., Ahmed, M., Novak, D., 2018. Identifying the causes of drivers' negative states using driver characteristics, vehicle kinematics and physiological measurements. Front. Neurosci 12, 568.

Darzi, A., McCrea, S., Novak, D., 2021. User experience comparison between five dynamic difficulty adjustment methods for an affective computer game. JMIR Serious Games 9 (2), e25771.

Darzi, A., Novak, D., 2021. Automated affect classification and task difficulty adaptation in a competitive scenario based on physiological linkage: an exploratory study. Int. J. Hum. Comput. Stud 153, 102673.

Dewan, M.A.A., Murshed, M., Lin, F., 2019. Engagement detection in online learning: a review. Smart Learning Environ. 6, 1. https://doi.org/10.1186/s40561-018-0080-z.

Eldenfria, A., Al-Samarraie, H., 2019. Towards an online continuous adaptation mechanism (OCAM) for enhanced engagement: an EEG study. Int. J. Hum. Comput. Interact 35 (20), 1960–1974. https://doi.org/10.1080/10447318.2019.1595303.

Ewing, K.C., Fairclough, S.H., Gilleade, K., 2016. Evaluation of an adaptive game that uses EEG measures validated during the design process as inputs to a biocybernetic loop. Front. Hum. Neurosci 10, 223. https://doi.org/10.3389/fnhum.2016.00223.

Fairclough, S.H., 2017. Physiological computing and intelligent adaptation. Emotions and Affect in Human Factors and Human-Computer Interaction, pp. 539–556. https://doi.org/10.1016/B978-0-12-801851-4.00020-3.

Fairclough, S.H., Karran, A.J., Gilleade, K., 2015. Classification accuracy from the perspective of the user: real-time interaction with physiological computing. In: Proceedings of the 33rd Annual Conference on Human Factors in Computing Systems (CHI '15), pp. 3029–3038.

Fairclough, S.H., Lotte, F., 2020. Grand challenges in neurotechnology and system neuroergonomics. Front. Neuroergon. 1, 602504 https://doi.org/10.3389/fnrgo.2020.602504.

Frees, E.W., 2004. Longitudinal and Panel data: Analysis and Applications in the Social Sciences. Cambridge University Press. https://doi.org/10.1017/CBO9780511790928.

Goršič, M., Darzi, A., Novak, D., 2017. Comparison of two difficulty adaptation strategies for competitive arm rehabilitation exercises. In: Proceedings of the 2017 IEEE International Conference on Rehabilitation Robotics, pp. 640–645.

Gosling, S.D., Rentfrow, P.J., Swann Jr, W.B., 2003. A very brief measure of the Big-Five personality domains. J Res Pers 37, 504–528. https://doi.org/10.1016/S0092-6566(03)00046-1.

Hargrove, L.J., Scheme, E.J., Englehart, K.B., Hudgins, B.S., 2010. Multiple binary classifications via linear discriminant analysis for improved controllability of a powered prosthesis. IEEE Trans. Neural Syst. Rehabil. Eng. 18 (1), 49–57. https://doi.org/10.1109/TNSRE.2009.2039590.

Hart, S., Staveland, L.E., 1988. Development of NASA-TLX (Task Load Index): results of empirical and theoretical research. In: Hancock, P.A., Meshkati, & N. (Eds.), Human Mental Workload. North Holland Press.

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

Liu, C., Agrawal, P., Sarkar, N., Chen, S., 2009. Dynamic difficulty adjustment in computer games through real-time anxiety-based affective feedback. Int. J. Hum. Comput. Interact 25 (6), 506–529. https://doi.org/10.1080/10447310902963944.

Liu, C., Conn, K., Sarkar, N., Stone, W., 2008. Online affect detection and robot behavior adaptation for intervention of children with autism. IEEE Trans. Rob. 24 (4), 883–896. http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=4598899.

McCrea, S.M., Geršak, G., Novak, D., 2017. Absolute and relative user perception of classification accuracy in an affective videogame. Interact. Comput 29 (2), 271–286.

Muthén, L.K., & Muthén, B.O. (2017). *Mplus User's Guide* (8th ed.). Muthén & Muthén.

Nacke, L.E., Kalyn, M., Lough, C., Mandryk, R.L., 2011. Biofeedback game design: using direct and indirect physiological control to enhance game interaction. In: CHI '11 Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 103–112.

Nagle, A., Wolf, P., Riener, R., 2016. Toward a system of customized video game mechanics based on player personality: relating the Big Five personality traits with difficulty adaptation in a first-person shooter game. Entertain. Comput 13, 10–24.

Novak, D., Mihelj, M., Munih, M., 2012. A survey of methods for data fusion and system adaptation using autonomic nervous system responses in physiological computing. Interact. Comput 24, 154–172. https://doi.org/10.1016/j.intcom.2012.04.003.

Novak, D., Nagle, A., Riener, R., 2014. Linking recognition accuracy and user experience in an affective feedback loop. IEEE Trans Affect Comput 5 (2), 168–172.

Novak, V.D., Hass, D., Hossain, M.S., Sowers, A.F., Clapp, J.D., & Clapp. (2022). *Effects of classification accuracy and adaptation magnitude in an affect-aware feedback loop for the Multi-Attribute Task Battery*. Zenodo. https://zenodo.org/record/7339589.

Picard, R.W., Vyzas, E., Healey, J., 2001. Toward machine emotional intelligence: analysis of affective physiological state. IEEE Trans. Pattern Anal Mach Intell 23 (10), 1175–1191.

Riek, L.D., 2012. Wizard of Oz studies in HRI: a systematic review and new reporting guidelines. J. Hum.-Robot Interac. 1 (1), 119–136.

Sadka, O., Antle, A., 2022. Interactive technologies for emotion regulation training: a scoping review. Int. J. Hum. Comput. Stud 168, 102906. https://doi.org/10.1016/J.IJHCS.2022.102906.

Samadiani, N., Huang, G., Cai, B., Luo, W., Chi, C.H., Xiang, Y., He, J., 2019. A review on automatic facial expression recognition systems assisted by multimodal sensor data. Sensors (Switzerland) 19 (8), 1863. https://doi.org/10.3390/s19081863.

Sanghavi, H., Zhang, Y., Jeon, M., 2023. Exploring the influence of driver affective state and auditory display urgency on takeover performance in semi-automated vehicles: experiment and modelling. Int. J. Hum. Comput. Stud 171, 102979. https://doi.org/10.1016/J.IJHCS.2022.102979.

Santiago-Espada, Y., Myer, R.R., Latorella, K.A., Comstock, J.R., 2011. The multi-attribute task battery II (MATB-II) software for human performance and workload research: a user's guide. NASA Technical Memorandum, pp. 2011–217164.

Shirzad, N., Van der Loos, H.F.M., 2016. Evaluating the user experience of exercising reaching motions with a robot that predicts desired movement difficulty. J. Mot. Behav 48 (1), 31–46.

Shu, L., Xie, J., Yang, M., Li, Z.Z., Li, Z.Z., Liao, D., Xu, X., Yang, X., 2018. A review of emotion recognition using physiological signals. Sensors (Switzerland) 18 (7), 2074.

Stephens-Fripp, B., Naghdy, F., Stirling, D., Naghdy, G., 2017. Automatic affect perception based on body gait and posture: a survey. Int. J. Soc. Robot 9 (5), 617–641. https://doi.org/10.1007/s12369-017-0427-6.

Ung, W.C., Meriaudeau, F., Tang, T.B., 2018. Optimizing mental workload by functional near-infrared spectroscopy based dynamic difficulty adjustment. Proc. Annual Int. Conference IEEE Engineer. Med. Biol. Society, EMBS. https://doi.org/10.1109/EMBC.2018.8512501.

Wilson, G.F., Russell, C.A., 2007. Performance enhancement in an uninhabited air vehicle task using psychophysiologically determined adaptive aiding. Hum. Factors 49 (6), 1005–1018. https://doi.org/10.1518/001872007X249875.

Xu, G., Gao, X., Pan, L., Chen, S., Wang, Q., Zhu, B., Li, J., 2018. Anxiety detection and training task adaptation in robot-assisted active stroke rehabilitation. Int. J. Adv. Rob. Syst. (6), 15. https://doi.org/10.1177/1729881418806433.