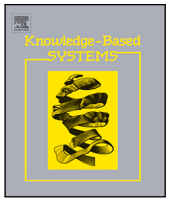
Knowledge-Based Systems 280 (2023) 111052

Contents lists available at ScienceDirect

Knowledge-Based Systems

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



Fusing ECG signals and IRT models for task difficulty prediction in computerised educational systems

Miguel Arevalillo-Herráez [a,*], Stamos Katsigiannis [b], Fehaid Alqahtani [c], Pablo Arnau-González [a]

[a] Departament d'Informatica, Universitat de Valencia, Burjassot, Valencia, 46100, Spain
[b] Department of Computer Science, Durham University, Durham University Upper Mountjoy Campus, Stockton Rd, Durham, DH1 3LE, United Kingdom
[c] Department of Computer Science, King Fahad Naval Academy, Al Jubail, 35512, Saudi Arabia

ARTICLE INFO

Keywords:
Task difficulty estimation
Adaptive learning
Dynamic difficulty adjustment

ABSTRACT

Accurately assessing task difficulty is a critical aspect to achieve adaptation in computer-based educational systems. In real-world scenarios, task difficulty estimation can be personalised for individuals by leveraging Item Response Theory (IRT) to analyse the collective performance of a group of students across various tasks. Additionally, recent studies have revealed the potential of inferring task difficulty through the analysis of physiological signals, such as electrocardiography (ECG). In this paper, we propose a novel hybrid approach that combines both methodologies to enhance task difficulty estimates, surpassing the individual performance of each method. The availability of non-intrusive techniques for capturing heart rate adds further value to the proposal, facilitating its potential integration into future computer-based educational systems. Experimental results on a dataset captured during two computerised English tests show that our proposed hybrid approach outperforms each individual method for the task of difficulty estimation.

1. Introduction

The adaptation of difficulty has been investigated in a wide range of domains. In the gaming industry, Dynamic Difficulty Adjustment (DDA) enables the fine-tuning of games to align with the player's desired flow state, ensuring that the game maintains optimal levels of motivation and challenge to enhance the overall gaming experience [1–3]. In the field of computer-assisted instruction, DDA has been widely applied to serve various purposes. One typical use has been to adaptively regulate task difficulty to maintain an optimal user performance level throughout a series of activities [4,5]. Another use has been to personalise the sequence of educational content to maximise learning gains over time [6,7]. Additionally, DDA has been employed to generate adaptive tests for assessment purposes [8] and issue recommendations [9], among other educational tasks.

In traditional one-to-one teaching situations, instructors are responsible for discerning the needs of learners and taking appropriate actions to guide the learning process. Nevertheless, in group learning settings or online remote learning, teachers may have limited contact with each learner. Intelligent Tutoring Systems (ITS) are computer systems that seek to resolve this issue by providing a tailored learning experience that is customised to the knowledge, needs, and abilities of each learner. Their advantage over traditional teaching and learning methods is their ability to provide automated knowledge tracing [10,11], as well as their capacity to adapt the learning process to match each student's unique requirements, resulting in a personalised learning experience [12,13]. This adaptation process usually focuses on offering learning activities that both challenge and motivate the student, and also on providing adequate support when the system detects that the learner is facing difficulties. To enable such adaptation capabilities, it is necessary to have estimation methods that can accurately determine a student's abilities and the level of difficulty associated with a particular task [14].

Item Response Theory (IRT) has proven successful in estimating students' abilities and task difficulties within computer-based learning environments, e.g. [15,16]. Recent attempts have also been made at predicting difficulty by using different types of physiological signals [17,18]. While some signals may require a certain level of intrusiveness, others can be easily obtained without any intrusion at all. This is the case with Electrocardiography (ECG) signals [19], which measure the electrical activity of the heart and can be easily captured by using typical wearables, such as activity wristbands, or by using camera-based methods, e.g. remote photoplethysmography (rPPG) [20].

IRT methods lead to a more accurate task difficulty estimation, despite making some assumptions that generally do not hold in learning

* Corresponding author.
E-mail addresses: miguel.arevalillo@uv.es (M. Arevalillo-Herráez), stamos.katsigiannis@durham.ac.uk (S. Katsigiannis), f-alqahtani@rsnf.gov.sa (F. Alqahtani), pablo.arnau@uv.es (P. Arnau-González).

https://doi.org/10.1016/j.knosys.2023.111052
Received 24 July 2023; Received in revised form 24 September 2023; Accepted 30 September 2023
Available online 7 October 2023

systems. In particular, they assume that the ability of the learner does not change across different activities and that all tasks are independent [16]. On the contrary, Heart Rate (HR) variations are subject-dependent and ECG-based predictions are able to capture a different point-of-view and offer a unique perspective with regard to difficulty measurements [18]. This leads to the hypothesis that combining IRT-based and ECG-based estimates could result in higher performance than using either method in isolation.

In this paper, we validate this hypothesis on a concrete ITS, using two computerised English-level tests that were designed to measure the learner's competence in the English language. Learners completed a first test, received tutoring using a video tutorial, and then took a second test while physiological signals were recorded. After answering each question, they were asked to report their perceived difficulty level on a 5-point Likert scale [21]. Three machine-learning models were then built with the objective of predicting self-reported question difficulty levels. The first model used features obtained from physiological signals, while the second model relied on IRT to estimate question difficulty without using physiological signals. The third model used late fusion to integrate the outputs of the first two models and generate a combined score, in an attempt to produce more reliable results. The findings indicate that the hybrid prediction method outperformed the other models on various metrics, including the area under the ROC (Receiver Operating Characteristic) curve (AUC), F1-score, and accuracy, demonstrating its efficiency for the task at hand.

The current work presents a strong contribution towards effectively assessing task difficulty in the context of computer-supported learning. This information can then be used to personalise instruction, by providing the most appropriate content and also by identifying the optimal timing to offer instructional scaffolding within the computer-based learning setting. The contributions and novelty of the proposed work can be summarised as follows: (i) We propose a hybrid approach for estimating task difficulty in a computerised learning environment, based on physiological signals and item response theory. (ii) We validate our proposed models on real data captured from students during test taking. (iii) We provide an extensive evaluation and analysis of the proposed and examined methods.

The remainder of the paper is organised in four sections. In Section 2, we provide background information about the problem and explain why this research is important and relevant. In Section 3, we aim to provide a comprehensive understanding of the datasets used in our experiments and the experimental setting, together with the proposed hybrid method for difficulty estimation. Section 4 focuses on the evaluation of the proposal, and presents an analysis of the results. Finally, Section 6 summarises the main findings of the study and suggests possible directions for future research.

2. Background and previous work

Progress in the field of ITS has provided alternatives to conventional teaching methods by reducing the need for extensive tutor involvement and allowing for prompt and customised feedback to learners. However, there is no universally accepted collection of features and approaches that define an ITS. The term "ITS" encompasses a wide range of techniques aimed at achieving a common objective: enhancing the learning process [18]. The general structure of ITS typically comprises four modules [22]: (i) The expert module, which encompasses the system's domain knowledge to be conveyed to the learners, as well as methods for examining the actions and behaviours of learners throughout their interaction with the system. (ii) The student module, which collects and updates information about the learner throughout the learning journey, such as submitted answers, behaviours, knowledge level, learning style, and more. (iii) The pedagogical (also known as instruction or tutor) module, responsible for identifying knowledge gaps and implementing specific strategies or teaching methods to compensate for these gaps and overcome learning difficulties. This module employs various methods, including adaptive feedback, hints, recommendations, and guiding the learner's path. (iv) The user interface, facilitating communication and interaction between the ITS and the learner.

The ability of ITS to adapt to the learners' needs and capabilities constitutes one of their most significant advantages. It is widely accepted that scaffolding strategies and adaptive difficulty adjustments have a positive impact on computerised learning systems, and help maintain the student's interest and achieve a convenient state of flow [23–25]. These scaffolding/adaptive strategies are typically based on a student model, which is conveniently used to present students with challenging problems, sequence activities in increasing order of difficulty, and/or offer hints or specific aids in difficult steps. Typically, adaptation techniques depend on estimates of both activity difficulty and student ability, and employ specialised algorithms to support the adaptation process.

The first approaches to difficulty assessment relied on relatively simple measures, which are generally applicable in most domains [16, 26]. The success rate is defined as the proportion of students who correctly solved the item, and can seamlessly be computed in most cases. The failure rate is generally used instead, because it aligns with the natural progression of greater difficulty corresponding to a higher failure rate. The average or median response time is another commonly used estimator. Due to their highly skewed nature, response times are better represented by the median as a measure of centrality rather than the mean. Another option is to use the mean of log-transformed times. Additional measures can be used in more constrained situations to supplement the aforementioned measures. For example, when multiple attempts are allowed, the number of attempts can be considered. When available, the number of hints requested and the number of edits required to construct an answer on a text editor can also be taken into account [26].

Similarly, the easiest method for estimating a student's ability is to rely on the same kind of statistical estimates, but computed across each student, rather than across each question/item. However, these simple measures fail to recognise that ability and difficulty are intertwined variables and do not consider the existing implicit relation between the two variables. Students who achieve the same proportion of correct answers are considered to have equivalent abilities, regardless of the difficulty of the problems they were able to solve correctly.

Item Response Theory (IRT) has wielded significant influence in psychological and educational measurement. Psychometric models within IRT assume that a person's answer to an item depends on the properties of both the person and the item [27]. The most basic model, known as the one-parameter Rasch model for dichotomous data (right/wrong answers), characterises the probability p of subject s correctly answering a specific question q as a logistic function of the difference between the estimated person's ability (β_s) and item's difficulty (δ_q):

$$p(y_{sq} = 1|\beta_s, \delta_q) = \frac{e^{\beta_s - \delta_q}}{1 + e^{\beta_s - \delta_q}} \tag{1}$$

where $y_{sq} = 1$ denotes that the answer provided by user s to question q is correct.

Using this model, the ability and difficulty parameters are mapped to a common scale, enabling the estimation of a success probability for any given item and learner. When $\beta_s = \delta_q$, the probability of a correct response is 0.5; if a person's ability level is greater than the difficulty level of a question, the likelihood of them answering it correctly is greater than 0.5; and conversely, if the difficulty level of a question exceeds the person's ability level, the likelihood of answering it correctly is less than 0.5.

Following their success in affect recognition tasks [28], the use of physiological signals has recently been explored for difficulty estimation. In particular, electrocardiography (ECG) and electromyography (EMG) signals were used in [17,19] as source data for a supervised classification method, in an attempt to estimate the self-perceived difficulty

level of the questions in real-time. To this end, several classifiers were trained using self-reported difficulty values as labels. The conducted experiments showed promising results, achieving an F1-score of 75.49% in the most favourable case when using ECG signals. In [29], enjoyment, valence, arousal, and perceived difficulty task difficulty were estimated from 5 different physiological responses, namely respiration, skin conductance, electrocardiogram, and 2 facial electromyograms. For perceived difficulty, the highest classification accuracies achieved were 84.3% for the two-class classification and 60.5% for the three-class classification. One intriguing aspect of these approaches is the seamless and implicit integration of the student's ability into the computation of the difficulty estimate, as signal variations will depend on it. This is particularly noteworthy given the highly subjective nature of the concept of difficulty, which may be influenced by a wide range of user-related traits.

IRT and physiological signal-based prediction methods use different input data and provide two very different views of the same problem. The Rasch model is only based on collected data regarding correct/incorrect answers. On the contrary, physiological signal-based estimates only consider information related to alterations detected on ECG and EMG signals to compute a difficulty estimate. Consequently, we hypothesise that merging the estimates generated by both approaches can result in enhanced outcomes, surpassing the performance of each individual modality. In addition, the current state of technology makes it possible to obtain non-invasive ECG measurements through the use of activity wristbands or rPPG techniques. As a result, it is now practical to employ this approach in real educational settings, without compromising the applicability of the resulting method.

3. Methodology

3.1. Data acquisition

To evaluate our hypothesis that combining IRT-based and ECG-based estimates could improve the performance of difficulty prediction within an educational setting, we first recruited a total of 28 individuals (18 male and 10 female) to participate in this study. All participants were international students at the University of the West of Scotland and/or lived in the local area (Paisley and Glasgow, Scotland, United Kingdom), and had familiarity with computer usage and a basic understanding of the English language. Their age was from 20 to 35 years old ($\mu_{\text{age}} = 24.1$ years, $\sigma_{\text{age}} = 4.4$ years). It must be noted that this work obtained approval from the Ethics Committee of the University of the West of Scotland based on a detailed description of the data collection method and anonymisation procedures for the data captured.

During the experimental sessions, ECG signals for each participant were recorded for the whole duration of the experiment. To minimise any discomfort or intrusion for the participants and reduce potential biases stemming from the presence of equipment, a portable and wireless measurement device with lightweight sensors was chosen to collect the ECG signals. In particular, a SHIMMER™ v2 wireless sensor was used to record the ECG at a 256 Hz sampling rate, with the four electrodes being positioned on the standard locations, on both lower ribs and clavicle. Additionally, a laptop computer was utilised to record the transmitted signals and monitor their quality.

3.2. Experimental protocol

We based our experimental protocol on [18] and extended it to fit our intended research questions. To this end, all participants were asked not to consume caffeine or drugs before the experiment, in order to avoid any potential effects on the physiological signals. In addition, the experimental setting was carefully designed to ensure a quiet and distraction-free environment. It took place in an office with no external noises or disturbances. Participants were first given an explanation of the experimental procedure and were offered the opportunity to ask questions. Then, they were asked to voluntarily sign a consent form and sit in front of a computer. A member of the research team supervised the attachment of the physiological sensors to the participants, either by attaching them directly or by guiding the participants in attaching them themselves, especially when the electrodes had to be placed on the skin beneath their clothing. To minimise motion artefacts in the recorded signals, participants were instructed to avoid excessive body and head movement during the experiment. The experiment then started once the correct signal transmission and acquisition had been confirmed.

The data acquisition experiment consisted of the three stages described below.

3.2.1. Stage 1

In the first stage, participants were required to take a computerised English language test. Participants were not restricted by any time limit per question and were instructed to indicate the perceived difficulty level of each question after providing each answer, as one of the following values: "Very Easy", "Easy", "Moderate", "Hard", or "Very Hard". The test consisted of 20 questions extracted from the Oxford Quick Placement Test (QPT) [30]. Designed to measure the English language proficiency of test takers and accurately place them into levels that align with the Common European Framework of Reference for languages (CEFR), the Oxford QPT consists of 40 questions of varying difficulty, related to 4 different tasks. Task 1 measures the test taker's ability to understand the meanings of phrases in a short text. Task 2 focuses on grammatical forms and requires test takers to complete a short gaped text by selecting one of three options. Task 3 tests test takers' knowledge of pragmatic meaning and contextual information, particularly in verbal phenomena where the communicative meaning differs from the literal meaning. Task 4 is designed to assess the test takers' comprehension of form and meaning. It involves a long passage with gaps that the test takers need to fill in with the correct answers and evaluates whether the test takers possess sufficient knowledge of grammar and vocabulary to accurately complete the gaps. Our test included 5 questions for each of these tasks. Specifically, we included all 5 questions for tasks 1 and 2, 5 questions from the 10 questions in task 3, and 5 more from the 20 questions in task 4. The selection of the questions for tasks 3 and 4 was done randomly, and the selected questions were all presented in the same order as they originally were in the Oxford QPT.

3.2.2. Stage 2 (Video tutorial)

During stage 2, participants watched a 26-minute English language video tutorial that did not require any interaction. The tutorial was 26.8 min long and consisted of: (a) examples of how to answer test questions similar to those in the Oxford QPT, including explanations justifying the correct answer, (b) instructional videos on various aspects of the English language, and (c) a motivational speech to encourage learning, titled "*Go ahead, make up new words!*" [31]. The purpose of this video tutorial was simply to have two different sessions separated in time, to be able to provide a more robust evaluation of the proposed method.

3.2.3. Stage 3

The third stage followed the same procedure as Stage 1 but used a different test that was built using a different version of the Oxford QPT that contained different test questions than the ones used in Stage 1.

3.2.4. Dataset preparation

The experiment concluded once all three stages were completed and feedback was provided. The average duration of the experiment across participants was 45.9 min, with a standard deviation of 4.8 min. To facilitate a more reliable experimental evaluation of our proposed method, the data collected in stages 1 and 3 were treated separately, resulting in the construction of two distinct datasets D_{pre} and D_{post}, with

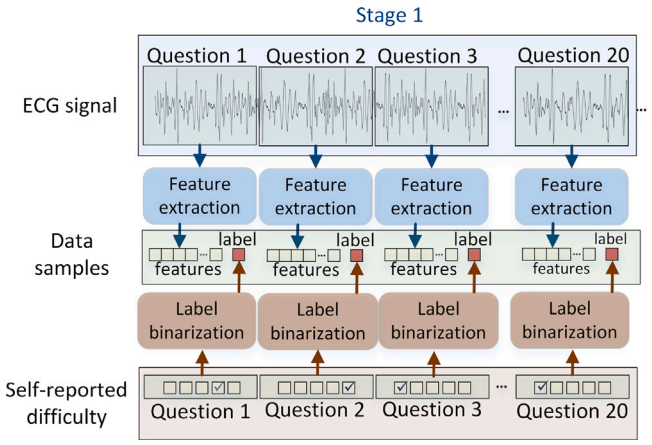


Fig. 1. Construction of the D_{pre} dataset used to estimate difficulty from the ECG signals.

pre referring to the English test conducted before the video tutorial, and *post* to the test conducted after the tutorial.

Following the protocol described in [18], after data acquisition, the examined problem was transformed to a binary classification one by converting the difficulty scores into a binary format by merging the samples rated as *Very easy* and *Easy* into the *Low* difficulty class, and the ones rated as *Hard* and *Very hard* into the *High* difficulty class. The samples corresponding to the *Moderate* difficulty level were excluded, leading to a total of 416 labelled samples in dataset $\mathcal{D}_{pre}$ and 435 samples in $\mathcal{D}_{post}$. It must be highlighted that the binary class labels were assigned by only considering the original user-reported difficulty scores, disregarding the correctness of the answer provided.

Note that the binarisation of difficulty scores has a two-fold effect in our setting. On one hand, it helps decrease the subjectivity associated with the adverb "very". On the other hand, it removes the intensity information that could further help learning algorithms to make more accurate predictions. Considering intensity would turn the problem into a regression one, introducing additional complexity due to the inherent subjectivity linked with grading difficulty. In this context, the utilisation of the two designated classes, namely "Low difficulty" and "High difficulty", along with the omission of samples labelled as "Moderate difficulty", serves not only to simplify the problem at hand but also to enhance the comparability of assessments made by various users.

3.3. Difficulty estimation from ECG signals

The acquired ECG signals were captured in a single continuous recording spanning the whole duration of the experiment for each participant. To reduce the effects of noise and artefacts, baseline wander was removed from the ECG signals [32], followed by a bandpass filter between 0.7–20 Hz. Then, for each dataset ($\mathcal{D}_{pre}$ and $\mathcal{D}_{post}$), the ECG signal recording for each participant was divided into 20 segments, each associated with one question. Each segment was then processed to compute the following spatial and spectral features using the Augsburg Biosignal Toolbox's (AuBT) [33] ECG feature extraction pipeline: minima, maxima, range ($maxima - minima$), mean, median and standard deviation of (a) the first-order difference of R-wave, P-wave, Q-peak, S-peak, T-wave indices, (b) the PQ, QS, ST indices difference, (c) the raw P, R, S signals subtracted by the mean of the raw PQST complexes' values, (d) the Heart Rate Variability (HRV), and (e) the HRV histogram; the number of intervals with latency > 50 ms from HRV divided by the total number of intervals; the total number of all intervals in the HRV histogram divided by the height of the HRV histogram; and the mean of the frequency spectrum of the HRV in the ranges [0 , 0.2], [0.2 , 0.4], [0.4 , 0.6] and [0.6 , 0.8] Hz. The resulting feature vectors for each segment were then annotated with the self-reported difficulty level assigned by the participant. The data processing performed in $\mathcal{D}_{pre}$ is illustrated in Fig. 1. $\mathcal{D}_{post}$ was processed in the same way, but using the data obtained during Stage 3 of the experiment.

The extracted features from the $\mathcal{D}_{pre}$ and the $\mathcal{D}_{post}$ datasets were then used in order to train machine learning models for each individual subject (participant) of this study for the task of difficulty prediction. To this end, four different classification algorithms were used to train five separate models for each participant, namely Linear Discriminant Analysis (LDA), Decision Tree (DT), Linear Support Vector Machine (LSVM) and k-Nearest Neighbour for $k = 1, 3$. To avoid overfitting and ensure a fair performance evaluation, difficulty estimates for each sample were produced by using the rest of the samples for the same participant as training data to build a subject-specific model.

3.4. Difficulty estimation using IRT

Difficulty estimates were also computed by using a one-parameter Rasch model for dichotomous data, using the girth v.0.8.0 python package implementation on the available data about the correctness of the answers provided by all subjects. For each sample, the entire $questions \times subjects$ (20 × 28) matrix M holding the correctness of the answers provided by all subjects in the corresponding dataset was used for the prediction, discarding the one associated with the particular sample. This is, for a subject s and a question q, the matrix element M_{qs} was designated as a missing value and predicted using the remaining matrix elements. This approach allowed the evaluation of the methods' performance by comparing the prediction to the original label.

The absolute difficulty parameter δ_q for each question q was estimated using Marginal Maximum Likelihood (MML), assuming $discrimination = 1$. Then, the ability for each subject s, β_s, was computed using Maximum Likelihood Estimation (MLE). Even though the package offers alternative methods to MLE and MML for estimating the abilities of subjects and the difficulty of questions, no significant differences were noted in the final results. The ability value allowed us to contextualise the absolute difficulty of each question based on the subject's skill level, resulting in a perceptual difficulty value pd_{qs}. Such contextualisation was achieved by adapting Eq. (1) to align with the difficulty concept, considering it as inversely proportional to this probability.

$$pd_{qs} = 1 - \frac{e^{\beta_s - \delta_q}}{1 + e^{\beta_s - \delta_q}} = \frac{1}{1 + e^{\beta_s - \delta_q}} \tag{2}$$

3.5. Difficulty estimation using the hybrid approach

The difficulty estimates obtained using the methods described in Sections 3.3 and 3.4 above were combined in an attempt to achieve a more reliable estimate of the perceived difficulty. To this end, we used a standard late fusion method, using a Linear Discriminant Analysis (LDA) classifier [34] to merge the scores obtained by using the ECG model built with all other samples from the same participant and the IRT model created using all other samples.

LDA assumes that variances are equal across classes and attempts to find a linear combination of features that best separates the classes. This is achieved by calculating statistical measures, such as class means and covariance matrices, to find a linear transformation that projects the data onto a lower-dimensional space in which the separation between classes is optimised. LDA aims to maximise the ratio of between-class variance to within-class variance, ensuring that classes are well-separated while minimising the variance within each class. This makes LDA effective for classification tasks, where the goal is to accurately categorise new data points into predefined classes based on their features. LDA is related to Principal Component Analysis (PCA), as they both look for linear combinations of the features that best explain the

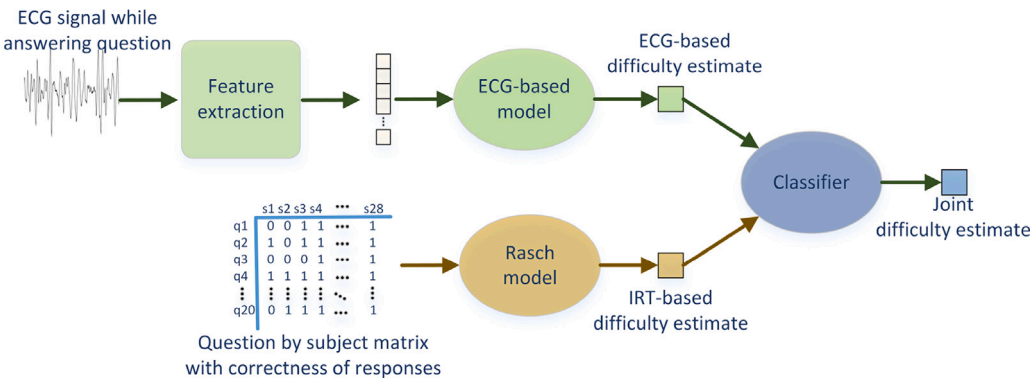


Fig. 2. Fusion of difficulty scores estimated by using ECG signals and IRT.

Table 1
Number of positive and negative samples in each dataset.

Dataset	Positive samples (High difficulty)	Negative samples (Low difficulty)	Total samples
$\mathcal{D}_{pre}$	314	102	416
$\mathcal{D}_{post}$	355	80	435

data. The main difference is that LDA is a supervised dimensionality reduction technique that also achieves the classification of the data simultaneously.

The choice of the LDA classifier was motivated by the simplicity of the fusion tasks, which is only acting on two dimensions; and the fact that this classifier does not require prior knowledge of the optimal parameter values, thus allowing for a more objective and unbiased assessment of our approach. Fig. 2 illustrates this approach, which takes the ECG signal and the correctness of responses by all users on the corresponding dataset as the input, in an attempt to leverage the information coming from both data sources.

4. Evaluation and results

4.1. Experimental setting

To fairly compare the quality of difficulty estimates using ECG signals, IRT, and the suggested hybrid approach, we have maintained a consistent experimental setup across all three methods. To make the most of our limited training data, we have used a leave-one-out cross-validation scheme. This technique involves sequentially designating each sample within the dataset as the test sample, while the rest of the samples are used for model training. The training and test process is repeated for every sample in the dataset, ensuring that each one is employed as a validation sample exactly once. Leave-one-out validation provides a rigorous evaluation of a model's generalisation ability, as it simulates the scenario where each sample is treated as entirely unseen during validation. Despite its computational demands, as the model must be trained and evaluated for each individual data point, our dataset's relatively modest size made this approach feasible.

4.2. Performance metrics

Table 1 presents the number of low and high-difficulty samples in each of the two datasets considered in this work. As it can be observed, low and high-difficulty judgements were unbalanced towards the high class, making accuracy (the proportion of correct predictions) an inadequate performance metric as the classifier could always achieve high accuracy by always predicting the majority class. In this case, the F1-score is a more appropriate metric that takes into account both precision and recall, providing a more reliable evaluation of classifier performance using a single number that takes the class imbalance into consideration. Precision measures the proportion of true positives among all predicted positives, while recall measures the proportion of true positives among all actual positives. The F1 score is the harmonic mean of precision and recall, taking into account both measures equally.

As we are not only concerned about the predicted class label but also about the predicted score as a difficulty estimate, the AUC-ROC (Area Under the Receiver Operating Characteristic curve) is also an interesting performance measure. In fact, the final label assignment should take into consideration the relative importance of false positives and false negatives, to set an appropriate threshold to classify the sample as either low or high difficulty. To put it in the context of real-world learning systems, if the difficulty score is used as a trigger to provide unsolicited assistance, false negatives would lead to instances where aid is not given when the user is experiencing challenges, potentially going unnoticed by the user. Conversely, false positives would involve offering help when it is unnecessary and was not requested, potentially yielding a perceptible adverse impact. AUC-ROC does not use a specific threshold to assign low or high labels. Instead, it considers the overall ranking of predicted scores for low and high-difficulty instances and provides a single threshold-independent summary score, which represents the overall ability of the classifier to assign higher scores to high-difficulty samples than to low-difficulty ones.

It must be noted that results reported in this work for all metrics correspond to the leave-one-out cross-validation scheme described in Section 4.1. They hence imply the use of 416 different models in the $\mathcal{D}_{pre}$ dataset and 435 in $\mathcal{D}_{post}$ (one per sample), providing a valuable ground for assessing the reliability of the tested models.

4.3. Difficulty estimation from ECG signals

Table 2 shows the average AUC, F1-score and accuracy results on the $\mathcal{D}_{pre}$ and $\mathcal{D}_{post}$ datasets, respectively. As it can be observed, and despite data scarcity (an average of 13.86 training samples for each subject-dependent model in $\mathcal{D}_{pre}$ and 14.54 in $\mathcal{D}_{post}$), the outcomes of this first experiment unambiguously indicate that the difficulty level of the question being asked has a significant impact on the ECG signal. The highest AUC was achieved using LDA to classify the ECG feature vector, reaching 0.744 and 0.746 for the $\mathcal{D}_{pre}$ and $\mathcal{D}_{post}$ datasets, respectively. On the contrary, when focusing on the F1-score, k-NN with $k = 3$ was the best-performing classifier in $\mathcal{D}_{pre}$ and the SVM performed best in $\mathcal{D}_{post}$, achieving values of 0.852 and 0.882, respectively. With regard to the accuracy, k-NN with $k = 3$ performed best on $\mathcal{D}_{pre}$, reaching a score of 0.776; and the SVM did best at $\mathcal{D}_{post}$, obtaining a 0.80 accuracy. Despite the differences in the best-performing algorithm for the F1-score and accuracy metrics, consistent results are obtained across the three different metrics considered, with only minor disparities between the rankings for each measure across the two different datasets.

Table 2
Performance of difficulty estimation using the ECG signals on the $\mathcal{D}_{pre}$ and $\mathcal{D}_{post}$ datasets.

Model	$\mathcal{D}_{pre}$			$\mathcal{D}_{post}$		
	Average AUC	Average F1-score	Average accuracy	Average AUC	Average F1-score	Average accuracy
LDA	**0.744**	0.827	0.743	**0.746**	0.875	0.795
Decision Tree	0.665	0.819	0.728	0.694	0.867	0.786
Linear SVM	0.740	0.851	0.762	0.710	**0.882**	**0.800**
k-NN ($k = 1$)	0.677	0.825	0.743	0.666	0.874	0.795
k-NN ($k = 3$)	0.742	**0.852**	**0.776**	0.708	0.876	0.793
Average	0.714	0.835	0.750	0.705	0.875	0.794
St. Dev.	0.039	0.016	0.019	0.029	0.005	0.005

Table 3
Performance of difficulty estimation using IRT on the $\mathcal{D}_{pre}$ and $\mathcal{D}_{post}$ datasets.

Dataset	Average AUC	Average F1-score	Average accuracy
$\mathcal{D}_{pre}$	0.833	0.864	0.796
$\mathcal{D}_{post}$	0.814	0.888	0.809

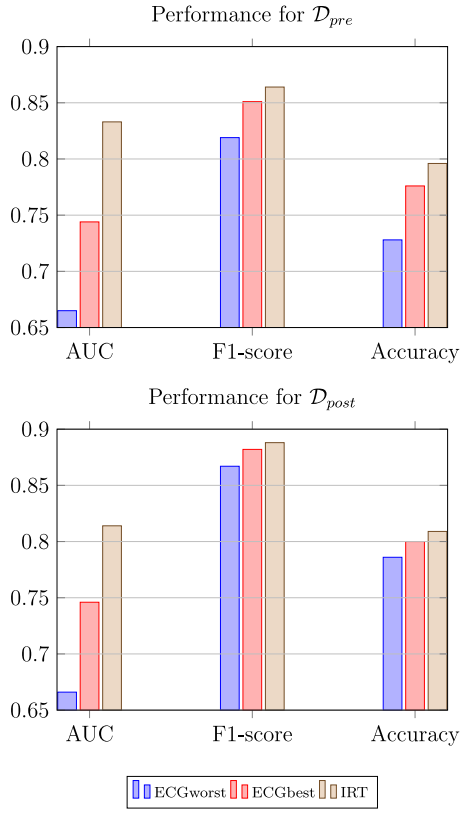


Fig. 3. Comparison of different performance metrics for difficulty estimation using ECG signals and IRT for the $\mathcal{D}_{pre}$ and $\mathcal{D}_{post}$ datasets. In the ECG case, both the top-performing (ECGbest) and lowest-performing (ECGworst) classifiers' results are presented.

4.4. Difficulty estimation using IRT

Table 3 shows the results when fitting a Rasch model on the question by subject matrix containing the correctness of the responses for all users in each particular dataset. Fig. 3 compares all metrics considered against those reported for the best- and worst-performing classifier when using ECG signals to estimate question difficulty. At the sight of the plots, and even taking the best classifier as a reference, IRT-based estimations seem more reliable than those based on the ECG signals. IRT achieved an average AUC value of 0.833 for the $\mathcal{D}_{pre}$ dataset, compared to the highest average AUC of 0.744 for the ECG-based approach. Performance was similar for the $\mathcal{D}_{post}$ dataset, with IRT achieving an average AUC of 0.814, compared to the highest average AUC of 0.746 for the ECG-based approach.

The IRT model in $\mathcal{D}_{pre}$ achieved an average F1-score of 0.864, surpassing the F1-score of the best-performing ECG model, which was 0.852. Similarly, the accuracy attained by the IRT model in $\mathcal{D}_{pre}$ was 0.796, outperforming the accuracy of the top-performing ECG model, which was 0.776. Consistently, the F1-score and accuracy outcomes in $\mathcal{D}_{post}$ were similar, albeit with slightly smaller performance gaps. The IRT model achieved an F1-score of 0.888, surpassing the 0.882 F1-score obtained by the best-performing ECG model. Additionally, the accuracy of the IRT model in $\mathcal{D}_{post}$ was 0.809, exceeding the 0.80 accuracy reported for the top-performing ECG model.

While the F1-scores and accuracy values for the IRT model are clearly superior to those of the ECG model, the most notable difference is observed in terms of AUC. One possible explanation for this could be that the threshold used to differentiate between the two classes is considerably distant from its optimal value. Given that the notions of low and high difficulty are highly subjective, the model's ranking ability is generally of greater interest than the classification label obtained after applying the threshold. Therefore, we should highlight the better behaviour observed regarding AUC in the IRT-based estimation, which clearly surpasses the one obtained by using ECG signals.

4.5. Difficulty estimation using the hybrid approach

Table 4 provides the values of the three performance metrics considered in this work when using an LDA classifier to combine the scores produced by the ECG- and IRT-based difficulty estimators, in the two datasets, $\mathcal{D}_{pre}$ and $\mathcal{D}_{post}$. For the $\mathcal{D}_{pre}$ dataset, the average AUC reached 0.856, the average F1-score reached 0.889, and the average accuracy reached 0.827 using LDA for the hybrid approach based on the k-NN ($k = 3$) ECG-based model. For the $\mathcal{D}_{pre}$ dataset, the best results were obtained when using the k-NN classifier with $k = 3$, although the results for the remaining classifiers were remarkably close across all metrics. When using the k-NN classifier, the average AUC reached 0.856, the average F1-score reached 0.889, and the average accuracy reached 0.827. For the $\mathcal{D}_{post}$ dataset, the average AUC reached 0.832 using LDA, the average F1-score reached 0.900 using k-NN ($k = 1$), and the average accuracy reached 0.832, also when using the k-NN ($k = 1$) classifier. Again, the performance obtained for the remaining classifiers was very close.

We should note that the hybrid approach outperforms the ECG-based and IRT-based approaches, as all reported values are higher than the ones reported in Table 2 for the ECG-based estimator, and also higher than the ones reported in Table 3 for the IRT-based estimator. Table 5 summarises the results by averaging values across all different classifiers for the IRT-based and hybrid approaches. These superior results can be visually observed by looking at Figs. 4 and 5, for the $\mathcal{D}_{pre}$ and $\mathcal{D}_{post}$ datasets, respectively. These plots depict a detailed representation of the results for the three different measures considered in the study, according to the classifier employed. They have been produced by combining the information provided by Tables 2, 3 and 4. Note that performance values for the IRT methods are constant across

Table 4
Performance of difficulty estimation using the hybrid approach on the $\mathcal{D}_{pre}$ and $\mathcal{D}_{post}$ datasets.

Model	$\mathcal{D}_{pre}$			$\mathcal{D}_{post}$		
	Average AUC	Average F1-score	Average accuracy	Average AUC	Average F1-score	Average accuracy
LDA	0.854	0.887	0.822	**0.832**	0.899	0.830
Decision Tree	0.835	0.883	0.817	0.831	0.898	0.828
Linear SVM	0.847	0.883	0.817	0.825	0.899	0.830
k-NN ($k = 1$)	0.842	0.886	0.822	0.826	**0.900**	**0.832**
k-NN ($k = 3$)	**0.856**	**0.889**	**0.827**	0.820	0.889	0.811
Average	0.847	0.886	0.821	0.827	0.897	0.826
St. Dev.	0.009	0.003	0.004	0.005	0.005	0.009

Table 5
Performance comparison of different approaches on the $\mathcal{D}_{pre}$ and $\mathcal{D}_{post}$ datasets. Values shown for ECG-based and hybrid methods are average values across all different classifiers.

Dataset	Method	Average AUC	Average F1-score	Average accuracy
$\mathcal{D}_{pre}$	Hybrid	**0.847**	**0.886**	**0.821**
	ECG-based	0.714	0.835	0.750
	IRT-based	0.833	0.864	0.796
$\mathcal{D}_{post}$	Hybrid	**0.827**	**0.897**	**0.826**
	ECG-based	0.705	0.875	0.794
	IRT-based	0.814	0.888	0.809

the classifier as they are obtained by using MLE in all cases. The plot clearly shows that IRT-based methods offer better performance than the ECG methods in all cases, but the combined score consistently surpasses the scores obtained by any of the ECG and IRT models when used independently. In addition, as can be seen from Table 4, the hybrid approach shows a notable insensitivity to the choice of classifier for the ECG-based model, as evidenced by the very low standard deviation reported for the three examined performance metrics across the five examined options. It is particularly relevant that the hybrid method outperforms the others even when using the least favourable classifier.

5. Limitations of current work

Despite the positive results reported in this paper, we should remark on certain limitations of the presented study. First, with regard to labelling, the binarisation of the difficulty scores helped simplify the classification problem, but also imposed restrictions on the evaluation of the approach, which did not take into consideration the original difficulty scores provided by the user. In addition, difficulty judgements have been considered independent, disregarding that the perceived difficulty of one question may influence the perceived difficulty of the next question.

Second, the proposed hybrid approach requires the use of physiological sensors, potentially impacting its practical viability in real-world scenarios. ECG signals can nowadays be captured non-intrusively using current technology, ranging from everyday wearables, such as wristbands, to more complex technologies like remote photoplethysmography (rPPG). However, the integration of such data acquisition devices into existing applications is not straightforward and still needs further development.

A third limitation that also relates to the practical applicability of the proposed method, involves ethical concerns. Whether ECG signals are captured using rPPG, wristbands, or alternative methods, they are highly personal and can provide intimate insights into an individual's health and well-being. Therefore, informed consent becomes a requirement, making learners fully aware of the purpose, risks, and benefits associated with their data being captured. In addition, it is crucial to ensure that proper measures are in place to protect the confidentiality and integrity of this sensitive data. Federated learning techniques may be able to address this issue and will be explored in our future work.

A fourth aspect relates to the boundaries of the work presented. While the study has concentrated on task difficulty estimation, it has not delved into its practical implications. Adaptive behaviour is highly application-dependent and has been intentionally excluded from the scope of the study.

Finally, a last issue relates to the scope of application of the proposed method. Due to the lack of standard databases that could be used to conduct the presented research, it became necessary to construct an experimental setting tailored to our objective. This setting was restricted to two computerised English tests, which may not be representative of other subjects or domains. While no evident barriers hinder the extension of these findings to different educational settings, the generalisation of the results to other domains should hence be taken with care.

6. Discussion and conclusion

Including scaffolding in learning activities has led to improved learning outcomes compared to using the same activities without scaffolding [25]. Difficulty estimation helps the development of scaffolding strategies by identifying situations in which learners struggle and require additional support. It also assists in selecting activities that offer an optimal level of challenge to students and helps the design of adequate learning trajectories.

In this work, we have introduced a difficulty estimation model that seamlessly integrates the use of ECG signals and IRT methods. ECG-based models are subject-based and therefore can capture the nuances and intricacies of each specific user. However, they need data from the same individual, hence limiting the amount of data that is available for training. On the contrary, IRT-based models are subject-independent and can make use of a larger amount of training data, although they may struggle to capture the specific intricacies that are inherent to each subject. By combining both approaches, we leverage the strengths of each method and achieve superior results. The subject-dependent ECG-based models provide personalised insights, while the user-independent IRT-based models offer a broader understanding of common patterns. This hybrid approach allows for a more comprehensive analysis that surpasses the capabilities of either method used separately and leads to more accurate estimates.

While additional physiological signals can be incorporated into the hybrid model to enhance the results, our primary focus was on the practical applicability of the presented approach in real-world settings. Hence, we specifically emphasised the use of data that can be seamlessly captured while maintaining a low level of intrusiveness. From this point of view, the data used by the IRT model can be captured seamlessly and transparently to the user. Regarding ECG signals, rPPG methods offer the capability to compute HR from video, replacing the need for more intrusive sensors such as the SHIMMER™ v2 wireless sensor used in this work. Moreover, the rapid advancements of activity wristbands and other personal devices suggest that alternative non-intrusive methods for measuring ECG may become available shortly. In contrast, other common physiological signals such as EMG or electroencephalography (EEG) require more invasive devices to accurately capture the required signals.

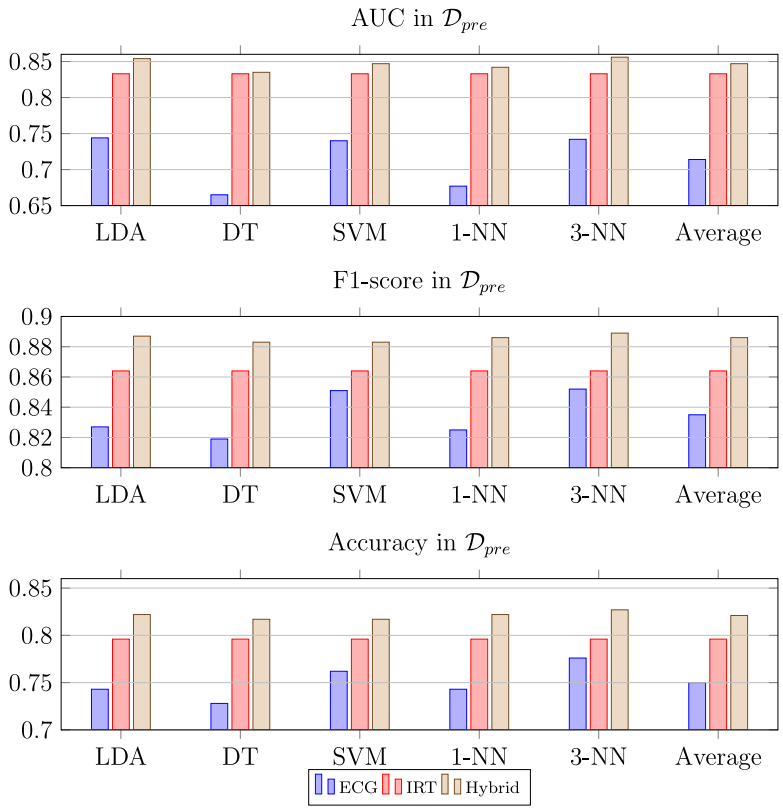


Fig. 4. Comparison of difficulty estimation using ECG signals, IRT and the hybrid approach, for the $\mathcal{D}_{pre}$ dataset, according to the classifier used. Average results across all classifiers are also reported on the right-most bars. The evaluation includes the three different measures considered in the study, namely AUC, F1-scores, and accuracy. The top row displays the AUC results, the middle row showcases F1-scores, and the bottom row illustrates accuracy results.

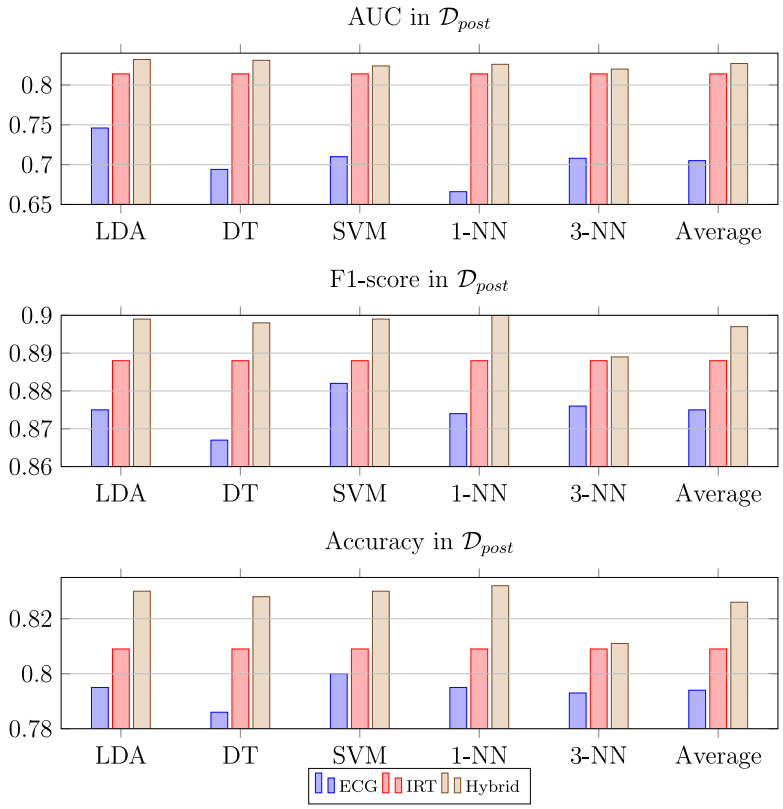


Fig. 5. Comparison of difficulty estimation using ECG signals, IRT and the hybrid approach, for the $\mathcal{D}_{post}$ dataset, according to the classifier used. Average results across all classifiers are also reported on the right-most bars. The evaluation includes the three different measures considered in the study, namely AUC, F1-scores, and accuracy. The top row displays the AUC results, the middle row showcases F1-scores, and the bottom row illustrates accuracy results.

We should also highlight the impact of the cold start problem on the applicability of the proposed method. This is a shared challenge to most recommender systems [35] and arises when a new user or activity is introduced to the system. In the case of an ECG-based model, there may not be an existing model that can be directly applied to the new user. Similarly, the IRT-based model requires initial data to estimate the student's ability β_s accurately, which is necessary for estimating task difficulty by using Eq. (2). Building an ECG-based subject-dependent model for new users requires explicitly asking them about the difficulty experienced with the first activities offered to them, and building a new model combining their responses with their ECG signals. When introducing new activities, task complexity [26] can initially be estimated by analysing its internal structure, serving as a proxy [36] for difficulty. For new users, an average ability β_s can initially be assumed, and further refined as additional data becomes available. Therefore, only when sufficient data become available, the proposed method allows for the inclusion of the individual's perception in the estimation process. On the positive side, it is worth noting that the accuracy of the estimation is expected to improve as new data is incorporated into the system. This new data will increase the size of the M_{qs} matrix and allow the IRT-based model to make more informed decisions, improving its estimation capabilities.

Future work will take several potential directions, primarily focused on addressing the limitations outlined in Section 5. First, this study has exclusively focused on task difficulty estimation and has not addressed other aspects of adaptive learning, such as personalised feedback or content recommendation. It is, therefore, a natural progression of this study to analyse the practical application of the proposed method to demonstrate the utility of estimating task difficulty within particular learning environments. We will also delve into the feasibility of employing alternative, less obtrusive, and more cost-effective capturing devices, while also scrutinising the impact of their precision on the hybrid method we have proposed. For example, rPPG exhibits optimal performance under controlled illumination conditions. Nonetheless, in a typical computer learning scenario, sudden illumination changes are infrequent and can be identified, allowing us to discard measurements taken during these occurrences.

It should also be noticed that the present study has been limited to a specific type of learning activity (computerised English tests). Further work will be required to replicate the approach across different learning contexts and confirm the broad applicability of the findings. At the same time, we shall extend the study to other alternative physiological signals, and analyse their correlation with the difficulty level. More complex fusion approaches that simultaneously consider several sources of information may become feasible and improve the accuracy of the prediction. Finally, intensive work will be required to ease the integration of the approach in real learning settings. Ideally, we should face the construction of an off-the-shelf component that makes use of the available physiological signals to feed recommender systems with an estimated difficulty value. We must not underestimate the complexity of this undertaking, which could potentially be a project in its own right, demanding substantial effort and human resources.

CRediT authorship contribution statement

Miguel Arevalillo-Herráez: Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Software, Resources, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization. **Stamos Katsigiannis:** Writing – review & editing, Visualization, Validation, Supervision, Methodology, Investigation, Formal analysis, Data curation. **Fehaid Alqahtani:** Supervision, Resources. **Pablo Arnau-González:** Supervision.

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

The authors do not have permission to share data.

Acknowledgements

This research has been supported by project TED2021-129485B-C42, funded by MCIN/AEI/10.13039/501100011033 and the European Union "NextGenerationEU"/PRTR; project AICO/2021/019, funded by Valencian Regional Government (Spain); and Margarita Salas 2022–2024 Grant awarded by Universitat de València, Spain (Ref:#MS21-29), funded by Spanish Ministry of Science, Innovation and Universities, Spain through NextGenerationEU funds.

References

[1] M. Zohaib, et al., Dynamic difficulty adjustment (DDA) in computer games: A review, Adv. Hum.-Comput. Interact. 2018 (2018).
[2] J.C. Lopes, R.P. Lopes, A review of dynamic difficulty adjustment methods for serious games, in: Optimization, Learning Algorithms and Applications: Second International Conference, OL2A 2022, PÓVoa de Varzim, Portugal, October 24-25, 2022, Proceedings, Springer, 2023, pp. 144–159.
[3] A.J. Seyderhelm, K.L. Blackmore, How hard is it really? Assessing game-task difficulty through real-time measures of performance and cognitive load, Simul. Gam. 54 (3) (2023) 294–321.
[4] Y. Zhang, W.-B. Goh, Personalized task difficulty adaptation based on reinforcement learning, User Model. User-Adapted Interact. 31 (2021) 753–784.
[5] J. Papoušek, V. Stanislav, R. Pelánek, Impact of question difficulty on engagement and learning, in: Intelligent Tutoring Systems: 13th International Conference, ITS 2016, Zagreb, Croatia, June 7-10, 2016. Proceedings, Vol. 13, Springer, 2016, pp. 267–272.
[6] D. Shi, T. Wang, H. Xing, H. Xu, A learning path recommendation model based on a multidimensional knowledge graph framework for e-learning, Knowl.-Based Syst. 195 (2020) 105618.
[7] A. Segal, Y. Ben David, J.J. Williams, K. Gal, Y. Shalom, Combining difficulty ranking with multi-armed bandits to sequence educational content, in: Artificial Intelligence in Education: 19th International Conference, AIED 2018, London, UK, June 27–30, 2018, Proceedings, Part II, Vol. 19, Springer, 2018, pp. 317–321.
[8] E. Guzmán, R. Conejo, J.-L. Pérez-de-la Cruz, Improving student performance using self-assessment tests, IEEE Intell. Syst. 22 (4) (2007) 46–52.
[9] S. Wan, Z. Niu, An e-learning recommendation approach based on the self-organization of learning resource, Knowl.-Based Syst. 160 (2018) 71–87.
[10] Y. Zhao, H. Ma, W. Wang, W. Gao, F. Yang, X. He, Exploiting multiple question factors for knowledge tracing, Expert Syst. Appl. 223 (2023) 119786, http://dx.doi.org/10.1016/j.eswa.2023.119786.
[11] T. Wu, Q. Ling, Fusing hybrid attentive network with self-supervised dual-channel heterogeneous graph for knowledge tracing, Expert Syst. Appl. 225 (2023) 120212, http://dx.doi.org/10.1016/j.eswa.2023.120212.
[12] X. Mao, Z. Li, Agent based affective tutoring systems: A pilot study, Comput. Educ. 55 (1) (2010) 202–208.
[13] S. Kuyoro, G. Maminor, R. Kanu, O. Akande, The design and implementation of a computer based testing system, History 5 (2016) 6.
[14] R. Conejo, E. Guzmán, J.-L.P. de-la Cruz, B. Barros, An empirical study on the quantitative notion of task difficulty, Expert Syst. Appl. 41 (2) (2014) 594–606, http://dx.doi.org/10.1016/j.eswa.2013.07.084.
[15] S.K. Milligan, P. Griffin, Understanding learning and learning design in MOOCs: A measurement-based interpretation, J. Learn. Anal. 3 (2) (2016) 88–115.
[16] Y. Lee, Estimating student ability and problem difficulty using item response theory (IRT) and TrueSkill, Inf. Discov. Deliv. 47 (2) (2019) 67–75.
[17] F. Alqahtani, S. Katsigiannis, N. Ramzan, On the use of ECG and EMG signals for question difficulty level prediction in the context of Intelligent Tutoring Systems, in: 2019 IEEE 19th International Conference on Bioinformatics and Bioengineering, BIBE, IEEE, 2019, pp. 392–396.
[18] F. Alqahtani, S. Katsigiannis, N. Ramzan, Using wearable physiological sensors for affect-aware intelligent tutoring systems, IEEE Sens. J. 21 (3) (2020) 3366–3378.
[19] F. Alqahtani, S. Katsigiannis, N. Ramzan, ECG-based affective computing for difficulty level prediction in intelligent tutoring systems, in: 2019 UK/China Emerging Technologies, UCET, IEEE, 2019, pp. 1–4.
[20] L. Malasinghe, S. Katsigiannis, K. Dahal, N. Ramzan, A comparative study of common steps in video-based remote heart rate detection methods, Expert Syst. Appl. 207 (2022) 117867, http://dx.doi.org/10.1016/j.eswa.2022.117867.
[21] A. Joshi, S. Kale, S. Chandel, D.K. Pal, Likert scale: Explored and explained, Br. J. Appl. Sci. Technol. 7 (4) (2015) 396.

[22] E. Mousavinasab, N. Zarifsanaiey, S.R.N. Kalhori, M. Rakhshan, L. Keikha, M.G. Saeedi, Intelligent tutoring systems: A systematic review of characteristics, applications, and evaluation methods, Interact. Learn. Environ. (2018) 1–22, http://dx.doi.org/10.1080/10494820.2018.1558257.

[23] S. Sampayo-Vargas, C.J. Cope, Z. He, G.J. Byrne, The effectiveness of adaptive difficulty adjustments on students' motivation and learning in an educational computer game, Comput. Educ. 69 (2013) 452–462.

[24] R.R. Van Der Stuyf, Scaffolding as a teaching strategy, Adolescent Learn. Dev. 52 (3) (2002) 5–18.

[25] M.-Y. Chang, W. Tarng, F.-Y. Shin, The effectiveness of scaffolding in a web-based, adaptive learning system, Int. J. Web-Based Learn. Teach. Technol. (IJWLTT) 4 (1) (2009) 1–15.

[26] R. Pelánek, T. Effenberger, J. Čechák, Complexity and difficulty of items in learning systems, Int. J. Artif. Intell. Educ. 32 (1) (2022) 196–232.

[27] I. Pandarova, T. Schmidt, J. Hartig, A. Boubekki, R.D. Jones, U. Brefeld, Predicting the difficulty of exercise items for dynamic difficulty adaptation in adaptive language tutoring, Int. J. Artif. Intell. Educ. 29 (2019) 342–367.

[28] S. Koelstra, C. Muhl, M. Soleymani, J.-S. Lee, A. Yazdani, T. Ebrahimi, T. Pun, A. Nijholt, I. Patras, Deap: A database for emotion analysis; using physiological signals, IEEE Trans. Affective Comput. 3 (1) (2011) 18–31.

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

[30] Q.O.P. Test, Quick Placement Test, Oxford University Press and the University of Cambridge Local, London, 2001.

[31] E. McKean, Go ahead, make up new words!, 2014, URL https://www.youtube.com/watch?v=pMUv6UWkuWw.

[32] N. Kannathal, U.R. Acharya, K.P. Joseph, L.C. Min, J.S. Suri, Analysis of electrocardiograms, in: Advances in Cardiac Signal Processing, Springer, 2007, pp. 55–82.

[33] J. Wagner, Augsburg Biosignal Toolbox (Aubt), University of Augsburg, 2005.

[34] G.J. McLachlan, Discriminant Analysis and Statistical Pattern Recognition, John Wiley & Sons, 2005.

[35] J. Feng, Z. Xia, X. Feng, J. Peng, RBPR: A hybrid model for the new user cold start problem in recommender systems, Knowl.-Based Syst. 214 (2021) 106732, http://dx.doi.org/10.1016/j.knosys.2020.106732.

[36] T. Effenberger, J. Čechák, R. Pelánek, Measuring difficulty of introductory programming tasks, in: Proceedings of the Sixth (2019) ACM Conference on Learning@ Scale, 2019, pp. 1–4.