Shin *et al.*
Large-scale Assessments in Education (2024) 12:6
https://doi.org/10.1186/s40536-024-00194-y

Large-scale Assessments in Education

RESEARCH Open Access

An engagement-aware predictive model to evaluate problem-solving performance from the study of adult skills' (PIAAC 2012) process data

Jinnie Shin[1*], Bowen Wang[1], Wallace N. Pinto Junior[1] and Mark J. Gierl[2]

*Correspondence: jinnie.shin@coe.ufl.edu

[1] Research and Evaluation Methodology, College of Education, University of Florida, Gainesville, FL, USA
[2] Department of Educational Psychology, University of Alberta, Edmonton, Canada

Abstract

The benefits of incorporating process information in a large-scale assessment with the complex micro-level evidence from the examinees (i.e., process log data) are well documented in the research across large-scale assessments and learning analytics. This study introduces a deep-learning-based approach to predictive modeling of the examinee's performance in sequential, interactive problem-solving tasks from a large-scale assessment of adults' educational competencies. The current methods disambiguate problem-solving behaviors using network analysis to inform the examinee's performance in a series of problem-solving tasks. The unique contribution of this framework lies in the introduction of an "effort-aware" system. The system considers the information regarding the examinee's task-engagement level to accurately predict their task performance. The study demonstrates the potential to introduce a high-performing deep learning model to learning analytics and examinee performance modeling in a large-scale problem-solving task environment collected from the OECD Programme for the International Assessment of Adult Competencies (PIAAC 2012) test in multiple countries, including the United States, South Korea, and the United Kingdom. Our findings indicated a close relationship between the examinee's engagement level and their problem-solving skills as well as the importance of modeling them together to have a better measure of students' problem-solving performance.

Keywords: PIAAC, PS-TRE, Long-short-term-memory (LSTM), Attention, Engagement, Problem-solving performance, Log information

Large-scale digital assessment in an interactive online environment is designed to evaluate examinees' thinking and problem-solving skills (Van Laar et al., 2017). An increasing number of large-scale assessments, such as the Programme for International Assessment of Adult Competencies (PIAAC), the Programme for International Student Assessment (PISA), and the Trends in International Mathematics and Science Study (TIMSS), have recently introduced more innovative test solutions with novel item formats to assess problem-solving or collaborative problem-solving performance (e.g., Barber et al., 2015;

Mullis et al., 2021). For example, PIAAC is an international assessment which was the first fully computer-based large-scale assessment in education and the first to provide public anonymized log file information widely.[1] PIAAC's problem-solving assessment in a technology-rich environment (PS-TRE hereafter) is designed to assess the adult examinee's ability to use "digital technology, communication tools and networks to acquire and evaluate information, communicate with others and perform practical tasks" (Rouet et al., 2009, p.9; (OECD), 2012, p. 47). In this test, examinees are provided with varying types of problem-solving tasks that embed authentic real-life scenarios.

These non-traditional, interactive, digital problem-solving items encourage examinees to demonstrate their authentic skill sets using their responses and the traces of activities associated with solving the task (Jiang et al., 2021). The traces stored as metadata of examinees' interactions are the process log data or click-stream information. The process log data provides insights into the examinee's behavior that are not easily disambiguated with the response data alone, especially in many non-traditional and interactive large-scale assessments. The process log information uncovers more individualized and diagnostic evidence about the examinees' latent abilities (Goldhammer et al., 2014; He & von Davier, 2015; Scherer et al., 2015; Wang et al., 2021) which enhances the reliability and validity evidence of the assessments (Kroehne & Goldhammer, 2018; Ramalingam & Adams, 2018), and identifies the examinees who are depicting anomalous behaviors (Lundgren & Eklöf, 2020). For instance, Jiang et al., (2021) demonstrated how the process data gathered specifically from students' drag-and-drop actions in a large-scale digitally-based assessment environment could infer examinees' varying levels of cognitive and metacognitive processes, such as their problem-solving strategies.

Incorporating the process information in a large-scale assessment to achieve such goals requires several methodological and empirical considerations. First, the complex micro-level evidence from the examinees (i.e., process log data) needs to be analyzed to extract explainable and interpretable patterns that inform the examinee's latent abilities (e.g., problem-solving strategies, Polyak et al., 2017; von Davier, 2017). Second, the examinees' demonstration of knowledge and skills need to be modeled in the sequences of task levels to provide more generalizable implications compared to the item-level results (Ai et al., 2019; Jiang et al., 2020; Liu et al., 2019a, 2019b; Wang et al., 2017). Third, careful consideration is required to evaluate the effect of students' sentiments or affect that may influence their performance, such as their task-disengagement behaviors (Wise, 2020).

With the recent wide introduction of machine learning and deep learning approaches in large-scale assessments and learning analytics, increasing attempts are being made to more effectively and efficiently analyze the process data from large-scale assessments. Hence, in this study, we propose a novel analytic framework where the examinee's complex and long traces of process log data are used to understand the problem-solving skills and performance. The present study is rooted in the fields of learning analytics and psychometrics. We combined multiple advanced computational methods, including social network analysis and deep neural network models. Our framework also models the examinee's task-engagement status for a more accurate representation of the

[1] https://www.oecd.org/skills/piaac/.

performance and skill demonstration in the series of interactive tasks. One research question is addressed to guide the study: *Does modeling the engagement levels with problem-solving skills improve the prediction performance of the LSTM model for items solved on a large-scale assessment?*

To describe how our research questions were addressed using the PIAAC's PS-TRE assessment, the subsequent sections focus on three primary topics. First, we present the construct measured by the PS-TRE test and its three core dimensions, thereby providing contextual information on the types of tasks our research aims to investigate and evaluate. Second, we offer an overview of the literature, concentrating on methodologies introduced to understand the PS-TRE construct, with a specific focus on recent studies that have utilized process data to model the tasks associated with this construct. Lastly, we provide an overview of how test engagement is currently modeled in various large-scale assessment settings, underscoring the importance of capturing test engagement in the PS-TRE.

Problem-solving tasks in PIAAC PS-TRE

The PIAAC's problem-solving assessment in a technology-rich environment (PS-TRE) is designed to assess the adult examinee's ability to use "digital technology, communication tools and networks to acquire and evaluate information, communicate with others and perform practical tasks" (OECD, 2012). Problem-solving usually means that people cannot solve problems through routine activities, which needs a complex hierarchy of cognitive skills and processes. Technology-rich environment indicates that some technologies (e.g., spreadsheets, Internet search, websites, email, social media, or their combination) are required to solve the task in assessment (Vanek, 2017).

The three core dimensions of PS-TRE include task/problem statements, technologies, and cognitive dimensions. These dimensions are closely connected because examinees rely on their choice of technologies to solve the problems, which requires their cognitive skills to successfully use the selected technology to solve the problem or accomplish the task. The examinees are provided with varying types of problem-solving tasks that embed authentic real-life scenarios based on intertwined dimensions of PS-TRE. A problem-solving task can be provided by connecting any domains in each core conceptual dimension as described in Appendix 1.

The interactions between these three core components create complex problem-solving tasks. Examinees are required to use a sequence of actions to correctly address these tasks, resulting in a substantial collection of process logs and clickstream information. The following section will explore the modeling of this extensive data, aiming to extract meaningful insights into the problem-solving strategies used by examinees during the assessment.

Modeling problem-solving strategies in PS-TRE with process data

Increasingly, studies have been conducted to introduce various computational and artificial intelligence-powered methods to effectively understand examinees' responses as well as the complex interaction process log information gathered in PIAAC'S PS-TRE. These studies often adopted clustering analysis (He, Liao, & Jiao, 2019), pattern mining analysis (Liao et al., 2019), graph modeling approaches (Ulitzsch et al., 2021), and

clickstream analysis with supervised learning models (Ulitzsch et al., 2022). For instance, He et al., (2019a, b) adopted the K-means algorithm (Lloyd, 1982) to cluster the behavioral patterns from one representative PS-TRE item based on features extracted from process data (i.e., unigrams by n-grams model, total response time, and the length of action sequences) to explore the relationship between behavioral patterns and proficiency estimates covaried by employment-based background. That is, more actions and longer response time tended to generate higher PS-TRE scores when getting incorrect answers. Their findings indicated process data tends to be more informative when items are not answered correctly.

To further investigate the impact of employment-based background, Liao et al., (2019) mapped the employment-based variables with action sequences in process data using the regression analysis and chi-square feature selection mode. They found that groups with different levels of employment-based background variables tended to generate distinctive characteristics regarding the action sequences to solve problems. However, it should be noted that the previous approaches (e.g., He et al., 2019a, b; Liao et al., 2019) only analyzed item-level timing data instead of time consumed between actions (i.e., action-level time), so the more detailed underlying cognitive processes due to time-stamped action sequences might be neglected.

Ulitzsch et al., (2021) proposed a two-step approach to analyze complete information contained in time-stamped action sequences for a deeper investigation of the behavioral processes underlying task completion. The researchers integrated tools from clickstream analyses and graph-modeled data clustering with psychometrics so that they can combine action sequences and action-level times into one analysis framework. In another study, enriching generic features extracted from sequence data by clickstream analysis, Ulitzsch et al., (2022) extracted features from time-stamped early action sequences (i.e., early-window clickstream data) and an extreme gradient boosting (XGBoost) classifier was used (Chen & Guestrin, 2016). Within the procedure, early-window datasets were created to train the model by getting rid of all afterward time-stamped actions (i.e., occurred after a given number of actions or a given amount of time from the sequences) thereby allowing the features taken from clickstreams to focus on the occurrence, frequency, and sequentiality of actions by adding features based on the amount of time consumed to carry out certain actions. Based on the clickstream analysis with a supervised learning model, Ulitzsch et al., (2022) investigated the early predictability of success or failure on problem-solving tasks before examinees complete the tasks and deepen the understanding of the trajectories of behavioral patterns in PS-TRE.

These studies demonstrated excellent potential to leverage our understanding with interpretable results to study different facets of students' knowledge and abilities. However, no study, to our knowledge, was introduced in the sequence of task-level, with the potential to consider the examinees' engagement status in the analysis of problem-solving knowledge modeling. Therefore, in the subsequent section, we will introduce how test-taking engagement has been defined in previous literature, along with the methodologies explored to investigate such constructs. Consequently, we will highlight the benefits and advantages of employing test-taking engagement as a simultaneous measure to effectively evaluate students' performance.

Engagement in knowledge modeling with problem-solving performance in large-scale assessment

Test-taking engagement is used to describe if the test taker remains engaged throughout a test, which is an underlying assumption of using all psychometrics models in practice (Wise, 2015, 2017). The term test-taking engagement also refers to the test-taking effort. Test disengagement was defined as providing or omitting responses to items with no adequate effort (Kuang & Sahin, 2023), indicated by rapid-guessing behavior (Schnipke, 1996; Wise & Kong, 2005) and item-skipping behavior. A lack of test-taking engagement is a major threat to the validity of test score interpretation even in good test design (Wise & DeMars, 2006), especially in low-stakes assessments such as PIAAC (Goldhammer et al., 2016).

Modeling test-taking engagement in problem-solving tasks resolves the potential validity threat (e.g., construct-irrelevant variance) that can confound the examinees' performance results (Braun et al., 2011; Goldhammer et al., 2016; Keslair, 2018; Wise, 2020). Information gathered from the examinee's response data in the task was commonly used to model their task-engagement level. Various item response theory (IRT)-based models incorporate students' engagement to predict people's latent traits (Deribo et al., 2021; Liu et al., 2019a, b; Wise & DeMars, 2006). For instance, Wise and DeMars, (2006) introduced the effort-moderate IRT (EM-IRT) model, where disengaged responses are treated as missing data and fit the engaged responses to a unidimensional IRT model. Response time was used to identify students' engagement in the EM-IRT model. More recently, studies explored the use of data gathered from the interactions, such as process log data, to evaluate examinees' test-taking effort and motivation (Lundgren & Eklöf, 2020; 2021).

The combination of response time and response behaviors was used as an "enhanced" method to detect examinees' disengagement (Sahin & Colvin, 2020). Within this approach, the response behaviors (e.g., keypresses, clicks, and clicking interactive tools) were collected from the process data (Kuang & Sahin, 2023). Sahin and Colvin, (2020) set up the threshold for the maximum number of response behaviors that suggest no or minimum engagement. However, they did not use statistical models to analyze the response behaviors from process data. A small number of studies have demonstrated the capacity to model the examinee's engagement and problem-solving performance from process data or a sequence of tasks (as well as at an individual task level). Since test engagement can be treated as a latent trait under response behaviors and deep learning approaches have the advantage of modeling process data or a sequence of tasks to capture examinees' response behaviors, it is worth investigating how to apply deep learning approaches (such as Long Short-Term Memory Networks) to detect test engagement.

Long short-term memory networks

Our study implements one of the variational models of recurrent neural network (RNN) models to effectively and accurately track students' problem-solving performance from a sequence of PS-TRE tasks. Unlike traditional neural network models, the RNN models introduce a simple loop structure in the hidden layer to consider a sequence or a history of input. In our study, we use one of the special variations of the

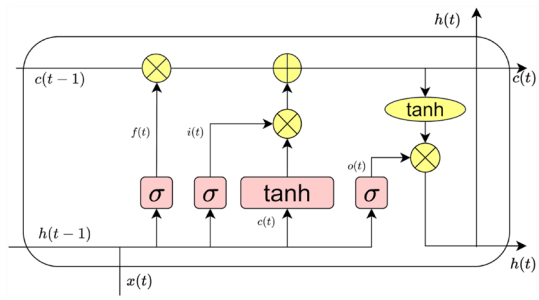


Fig. 1 A Conceptual Representation of an LSTM Memory Cell

RNN models, which is the Long Short-Term Memory (LSTM) network. The LSTM model consists of units called memory blocks. Each memory block consists of multiple gates—input, forget, and output gates—that control the flow of information.

Figure 1 provides an overview of an example LSTM memory cell structure. In our study, we use the memory cell to input, modify, extract, and communicate the deterministic information about examinee's problem-solving strategies and performance on a sequence of tasks, where t represents the task that the examinee is interacting with. Specifically, input data is determined based on n batch size with d features and h number of hidden layers, $x(t) \in R^{n\times d}$, and the hidden state of the previous task $h(t-1) \in R^{n\times h}$, indicating the final input data as $X^T = [h(t-1), x(t)]$. This input data is first provided to the forget gate $f(t) \in R^{n\times h}$, input gate $i(t) \in R^{n\times h}$, and an output gate $o(t) \in R^{n\times h}$. The forget gate governs the degree to which the information from the previous tasks is omitted from the cell state, the input gate governs how much new information about the examinee's problem-solving skills are inferred from the current task, and the output gate produces the output that will be communicated to the next cell state for the task, $t+1$.

The interim values after entering the gates are computed as below, where w_{xi}, $w_{xf}, w_{xo} \in R^{d\times h}$ and w_{hi}, $w_{hf}, w_{ho} \in R^{h\times h}$ represent weights of each gate, and b_i, $b_f, b_o \in R^{1\times h}$ represent bias of each gate, respectively. The input node $\widetilde{c}(t) \in R^{n\times h}$ is also computed similarity with the other gates, where the activation function of $tanh(x) = (e^x - -e^{-x})/(e^x + e^{-x})$ replaces the sigmoid function in the other three gates.

$$\begin{aligned} i(t) &= \sigma(x(t)w_{xi} + h(t-1)w_{hi} + b_i), \\ f(t) &= \sigma(x(t)w_{xf} + h(t-1)w_{hf} + b_f), \\ o(t) &= \sigma(x(t)w_{xo} + h(t-1)w_{ho} + b_o) \\ \tilde{c}(t) &= tanh(x(t)w_{xc} + h(t-1)w_{hc} + b_c) \end{aligned} \tag{1}$$

The memory cell outputs the internal state and the hidden state $h(t) \in [-1, 1]$. The hidden state at task t will concern the input, forget, and output gates by deciding the impact of the current memory to the next memory cell. The hidden state that is close to the value of 0 will minimize the current impact to the next cell while the value close to 1 will impact the internal state value of the next cell with no restriction. The memory cell updates the internal state $c(t)$ in the task t by gathering the information from the forget, input, and the previous cell state as follows:

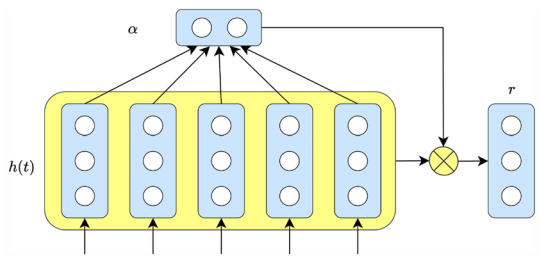


Fig. 2 A Conceptual Representation of the Attention Layer in LSTM

$$c(t) = f(t) \otimes c(t-1) \oplus i(t) \otimes \tilde{c}(t) \text{ and}$$
$$h(t) = o(t) \cdot tahn(c(t)). \tag{2}$$

Attention mechanism

The simple LSTM model can be limited in detecting which element provides the important aspect of information to determine examinees' problem-solving performance while accounting for their engagement level. Hence, we introduce an attention layer to explicitly model this information. Let $H \in R^{d \times t}$ represents the hidden layers derived from the memory cell of each problem-solving task t with an LSTM model with d hidden layers. The attention layer we use in the current finding is the global attention layer. The global attention layer represents the latent information extracted from the sequences of output from the encoder (i.e., input data is encoded using LSTM) in order to help decoders (i.e., output data is generated using LSTM) utilize global evidence related to examinees' problem-solving skills to output correct predictions. The dot-product attention computes the element-wise multiplication between the hidden states of encoder and decoder of task t, h_t and s_t with the attention weight $W = \{w_1, w_2, ..., w_n\}$, where the attention α is captured as follows:

$$\alpha = softmax(h_t^T W_a s_t) \tag{3}$$

Then, the final weighted representation of the hidden state is derived by combining the dot-product attention (α) and the hidden layer (H) as $r = H\alpha^T$. Using this information, we can represent the students' problem-solving performance as a combination of projection parameters W_p and W_r, are $h^* = sigmoid(W_p r + W_x h_n)$. The parameters W_p and W_x are learned during training (Rocktaschel et al., 2015). In our study, we use these projection parameters to visualize whether the attention layer is accurately capturing the examinee's problem-solving performance and engagement across a sequence of problem-solving tasks. The final univariate/multivariate outcome(s) (performance and engagement) of this process will be computed using h^*, as $y = softmax(W_s h^* + b_s)$, where W_s represents the output layer weights and b_s represents the output layer bias. This way we will be able to produce whether the student was engaged (=0), not engaged (=1), as well as the score category that the students acquired from the task as the final outcome of our model (see Fig. 2).

Using Long Short-Term Memory (LSTM) models to evaluate students' engagement and performance from process log data in the PS-TRE is a particularly effective approach

due to several key advantages of LSTMs. These neural networks are uniquely suited for handling sequential data, a core aspect of process log data, where the order and timing of actions are critical indicators of student engagement and performance. This allows us to evaluate students' performance and engagement effectively across multiple items and tasks, moving beyond analyzing the examinee's performance at an individual item level (e.g., Shin et al., 2022; Tang et al., 2016). LSTMs excel in capturing not just immediate dependencies but also long-term patterns in sequences, which is crucial in the PS-TRE context where early actions can influence later ones, or patterns of engagement may change over time. This indicates possibilities of capturing the information and storing the information from the examinee's process data at the very first task or the item they engage with, and utilizing their information to infer and predict their performance at the very last item they interact with.

The ability of LSTMs to learn complex patterns in sequential data is another significant advantage. They can handle variable-length sequences, a common characteristic in PS-TRE log data, ensuring consistent model performance across different data lengths (Hernández-Blanco et al., 2019). This aspect is vital, considering each examinee's interaction with the assessment varies in length and complexity. One of the standout features of LSTMs is their capacity for automatic feature extraction from raw sequential data. This is particularly beneficial for PS-TRE, where manually identifying relevant features from log data can be challenging. LSTMs can not only understand the context of each action within the broader sequence of events but also use this understanding to predict future behavior. This predictive ability is not only useful for analyzing past and present actions but also offers potential applications in real-time scenarios, such as adaptive testing or personalized learning interventions. Furthermore, LSTMs are robust to noise and irregularities in data, which are common in log files due to varied user behaviors and system inconsistencies (e.g., Fei & Yeung, 2015). Their capability to generalize from training data to unseen test data is vital for deploying models in different assessment environments.

Hence, the LSTM's proficiency in processing sequential data, its capability to detect and learn relevant features, and its robustness against data irregularities make it an appropriate choice for modeling the dynamics of student engagement and performance in PS-TRE. By leveraging the rich, time-ordered data in process logs, LSTMs provide deep insights crucial for educational assessments and learning analytics.

Data and materials

Data

We used the data collected from the first round of the OECD PIAAC Main Study, which was conducted from August 2011 to November 2012, involved 24 countries/economies, and was the first computer-based large-scale assessment to provide public anonymized log file data.[2] Our investigation focused on the cognitive domain of PS-TRE. A total of 14 tasks were dichotomously or polytomously scored (five 3-point, one 2-point, and 8 dichotomously scored items) (OECD, 2016). We analyzed the data collected from the

[2] https://search.gesis.org/research_data/ZA6712

Table 1 Demographic Information of the three Countries/Datasets of the Current Study

	PS-TRE Booklet *PS1*			
	Total N	**Age**	**Male (%)**	**Female e (%)**
United States	1,329	[a]	45.60	54.40
South Korea	1,434	34.98 (12.16)	46.93	53.06
United Kingdom	2,358	39.69 (13.29)	41.30	58.70

[a] Only available in the U.S. restricted-use files

United States (4131 units[3]), South Korea (7024 units), and the United Kingdom (7250 units). The log file of the PS-TRE tasks contained various information including the environment from which the event was issued (within the stimulus, outside of the stimulus), the event type, timestamps, and a detailed description of the event. In this proposed study, we experimented with the items included in one booklet (PS1) to demonstrate the prediction capacity of our proposed analytics framework (see Table 1).

Binary task engagement level

The method of T-disengagement (Goldhammer et al., 2016) was used to label test takers' engagement by response time as part of the training set. The term "T-disengagement" (OECD, 2019) describes a situation where examinees spend less time on a PIAAC task than a task-specific threshold. The approach to computing this item-specific threshold is based on the relationship between the probability of giving correct answers and the time spent on the item (Goldhammer et al., 2016). The underlying idea of this approach is that disengaged examinees tend to be less accurate than engaged examinees (Wise, 2017). The computation procedure first determined the time threshold t, it is necessary to compute the probability of getting a task correct on time t. The observations with a time on task between t and t + 10(s) are used. Then, the probability of correctness is modeled as a linear function of time if the number of the observations is enough (e.g., > 200). Last, the task-specific time threshold is determined as the smallest t for which the estimated probability of correctness is higher than 10%. The T-disengagement value was used in our study to create an engagement indicator, labeling test-takers' engagement based on response time as part of the training set. If an examinee spends less time on a task than the task-specific threshold, they are labeled as a disengaged examinee. Otherwise, they are considered engaged. Using the threshold calculated for each item in PS-TRE, we generated a binary outcome variable representing each examinee's engagement status.

Methods

Figure 3 provides a conceptual representation of our analytic model. Our analytic framework is based on a specific neural network model called the Long Short-Term Memory networks (LSTM; Hochreiter & Schmidhuber, 1997). The LSTM model takes a sequence of actions from the examinees which was captured while they were navigating through each item.

[3] According to https://search.gesis.org/research_data/ZA6712.

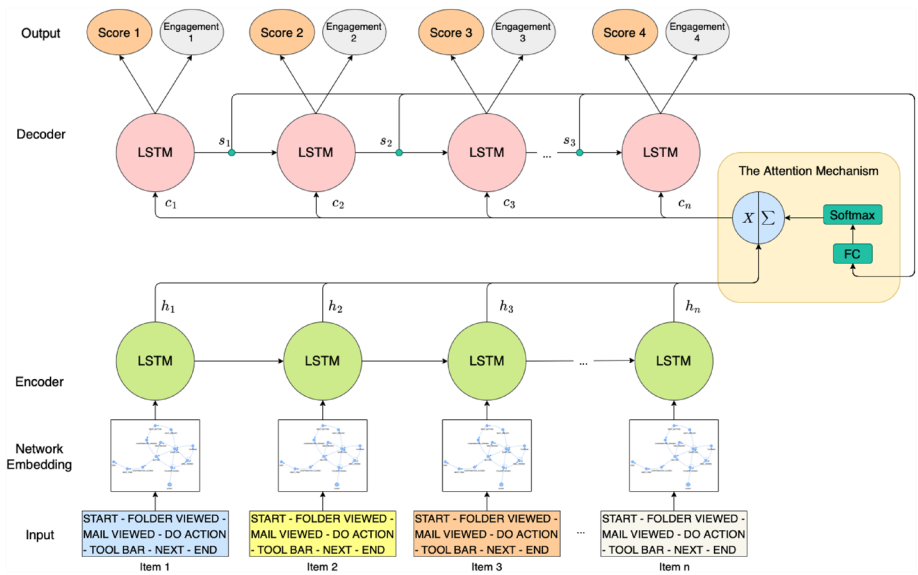


Fig. 3 A Conceptual Representation of the Effort-Aware Attention-LSTM Model

The first layer of the model focused on converting the input sequences of actions from the process log data into a directed graph, where a node represents an activity in an item and the edges represent the connectivity between the two actions. The edges are weighted by the total amount of time between the two actions. Then, the overall task-navigating process of the examinees was summarized using network statistics. Network statistics summarize the interactions present in the network. Our analysis adopted five network statistics. This method includes five key network statistics: centralization, density, flow hierarchy, shortest path, and total number of nodes, each contributing to a comprehensive understanding of the interactions within the network. This approach aligns with recent trends in educational data mining, where network analysis is increasingly applied to understand learning processes (Salles et al., 2020; Zhu et al., 2016).

Converting process log data into a directed graph in the first layer in LSTM for predictive modeling is a strategic decision that offers numerous benefits, particularly in the context of assessing complex sequential data like that found in PS-TRE. This conversion allows for a structured representation of the data, where each node in the graph represents an individual action or activity, and directed edges signify the sequence and transition between these actions. Importantly, by weighting these edges with the time elapsed between actions, the graph effectively captures the temporal dynamics integral to understanding examinee engagement and problem-solving processes.

This graph-based approach significantly enhances the analysis of sequential interactions among different actions (Zhu et al., 2016). It provides a more nuanced perspective on how examinees approach and navigate through tasks, revealing patterns and strategies in their problem-solving process. By employing network analysis techniques, such as evaluating centralization, density, flow hierarchy, shortest path, and the total number of nodes, the model can delve deeper into the complexity and

efficiency of examinees' approaches. Additionally, the directed graph structure is highly conducive to advanced machine learning techniques, such as those used in LSTM models, facilitating more accurate predictions and classifications based on the patterns identified in the graph (e.g., Zeng et al., 2021; Zhang & Guo, 2020). Beyond the analytical advantages, this representation also aids in the interpretability and visualization of the data, making it more accessible for educators and researchers to understand and visualize the problem-solving process. Moreover, this method's flexibility and scalability make it adaptable to various assessment scenarios, capable of accommodating different types of actions and interactions (Hanga et al., 2020). Overall, this first layer's approach of transforming log data into a directed graph lays a robust foundation for subsequent, in-depth analysis, capitalizing on the strengths of network analysis and machine learning to provide insightful interpretations of examinee behavior.

The encoder and decoder then summarized the network statistics and map them into the prediction outcomes. The encoder summarizes the input and represents it as an interim representation called internal state vectors. The decoder, on the other hand, generates sequences of output using the internal state vectors from the encoder as an input. In our study, we presented two variations of models that differ in the type and the number of outputs associated with the input. The first model (Attention-LSTM) only concerns the association between students' process activities (log information) and their performance outcome (i.e., categorical scores) in each task. The second model (Effort-Aware LSTM) additionally models the associations between students' process activities with their task engagement level to reduce any effects stemming from the low-stakes characteristics of the current dataset. In summary, the second model is designed to produce output regarding students' performance scores simultaneously with their task engagement level for each task.

In order to increase the interpretability of the model decisions (i.e., whether the model is correctly stipulating the information related to students' latent ability level), we included an attention mechanism. The global attention layer represents the latent information extracted from the sequences of output from the encoder in order to help decoders utilize global evidence related to examinees' problem-solving skills (Model 1) and problem-solving skills with engagement level (Model 2).

Evaluation

A two-step evaluation process was used. First, the two variations of the model were compared based on the overall and item (or task)-specific performance score prediction accuracies. In the first step of our evaluation process, we compared the two variations of the LSTM model based on their ability to predict overall and item-specific performance scores. To ensure a comprehensive assessment, we employed three evaluation metrics: accuracy, F1-score, and the area under the Receiver Operating Characteristic (ROC) curve. These metrics were chosen for their ability to provide a balanced view of the model's predictive performance, considering aspects like the balance between sensitivity and specificity (ROC curve) and the harmonic mean of precision and recall (F1-score). The final evaluation metrics were derived from the average results obtained through

Table 2 Experiment Results 1–Overall Average Prediction Performance

[a]DV	Attention-LSTM			Effort-Aware Attention-LSTM					
	Performance			Performance			Engagement		
	ACC	AUC	F1	ACC	AUC	F1	ACC	AUC	F1
US	0.779 (0.080)	0.751 (0.101)	0.824 (0.098)	0.802 (0.113)	0.821 (0.110)	0.875 (0.106)	0.872 (0.050)	0.878 (0.052)	0.937 (0.050)
SK	0.751 (0.091)	0.700 (0.102)	0.828 (0.110)	0.859 (0.115)	0.824 (0.120)	0.886 (0.108)	0.845 (0.017)	0.857 (0.020)	0.881 (0.048)
GB	0.778 (0.087)	0.771 (0.100)	0.816 (0.108)	0.883 (0.103)	0.835 (0.098)	0.900 (0.097)	0.844 (0.082)	0.858 (0.068)	0.920 (0.098)

US: United States; SK: South Korea; GB: United Kingdom

[a] DV: Dependent Variable; The performance and engagement level is simultaneously predicted in the second model

Table 3 Experiment Results 2–Task (Item)-level Average Prediction Performance

[a]DV	Item	Attention-LSTM		Effort-Aware Attention-LSTM			
		Performance		Performance		Engagement	
		ACC	F1	ACC	F1	ACC	F1
United States	1	0.774	0.861	0.802	0.857	0.893	0.971
	2	0.861	0.848	0.764	0.830	0.804	0.879
	3	0.792	0.824	0.720	0.785	0.865	0.897
	4	0.603	0.813	0.842	0.921	0.879	0.998
	5	0.723	0.773	0.976	0.981	0.949	0.938
South Korea	1	0.844	0.936	0.842	0.934	0.869	0.892
	2	0.833	0.815	0.820	0.817	0.821	0.833
	3	0.827	0.813	0.858	0.733	0.864	0.841
	4	0.642	0.824	0.824	0.952	0.860	0.885
	5	0.53	0.752	0.758	0.992	0.869	0.952
United Kingdom	1	0.823	0.884	0.812	0.887	0.866	0.975
	2	0.860	0.839	0.837	0.860	0.824	0.896
	3	0.834	0.841	0.845	0.861	0.769	0.759
	4	0.553	0.806	0.789	0.899	0.879	0.998
	5	0.686	0.711	0.891	0.991	0.952	0.970

[a] DV: Dependent Variable; The performance and engagement level is simultaneously predicted in the second model

threefold cross-validation. This cross-validation approach adds rigor to our evaluation, ensuring that the performance metrics are robust and not overly fitted to a specific partition of the data.

The second step involved conducting a Principal Component Analysis (PCA) on the final attention layer of our engagement-aware model (e.g., Chen et al., 2018). Applying PCA to the last attention layer of an LSTM model, which handles complex datasets related to student engagement and performance, offers significant benefits. Firstly, PCA is instrumental in reducing the dimensionality of high-dimensional outputs generated by the attention layer. This reduction is crucial, as it retains essential patterns and

variances while uncovering underlying latent associations. The ability of PCA to reveal latent relationships within the attention layer's output is particularly valuable. It exposes underlying structures that might not be immediately evident, providing deeper insights into how the model processes and combines various aspects of the input data (e.g., Qiao & Li, 2020; Zhang et al., 2020). Moreover, PCA helps validate the focus of the attention mechanism, ensuring that it aligns with features pertinent to the task. This validation is essential for confirming that the model adheres to theoretical and empirical expectations, ensuring that the predictive model is focusing on and depending on the 'adequate' source of information for the decision-making process (Terrin et al., 2003).

Results

Tables 2 and 3 provide the overall performance results of the two variations of the models proposed in this study. The results showed that the Effort-Aware Attention-LSTM model could achieve improved performance in predicting student performance scores in all three-evaluation metrics across all three countries. Our first model (Attention-LSTM) produced f1-scores close to 0.82, ROC of 0.70–0.75, and accuracy of 0.75–0.78 across all three countries. The second model (Engagement-Aware Attention-LSTM) produced f1-scores close to 0.88–0.90, ROC of 0.82–0.84, and accuracy of 0.80–0.88. The prediction performance on the examinee's engagement level produced f1-scores close to f1-scores 0.92–0.94, ROC of 0.86 to 0.88, and accuracy of 0.84–0.87. In summary, an improvement in the problem-solving performance prediction was observed in the second model.

For individual tasks (see Table 3), similar patterns were identified across the three countries where engagement-aware models acquired slightly improved performance results compared to the other model. The model results also demonstrated that the engagement-aware model could predict the engagement and disengagement level of the participants across all five tasks with high performance accuracies. Specifically, the improvement in prediction accuracy was the highest in Task 5 where F1-score improved by + 0.21 to + 0.28, and accuracy improved by + 0.20 to + 0.23.

Attention-layer visualization: engagement and performance latent variables

Appendix 2 provides visualizations of the attention layer from the engagement-aware model for each problem-solving task with the U.S. participant data set. The principal component analysis results visualized the potential underlying components that our attention mechanism captured to make correct decisions regarding students' performance results. The results showed that the interim output of the attention layer could be systematically explained by the two components which aligned with the problem-solving performance skill level with a relatively small variance explained by the second component, engagement level. The two components accounted for 75.5% and 14.4% of the

Table 4 The Pearson Correlation Coefficients between the Principal Components and Engagement and Performance

	PC1	PC2
Task 1		
Engagement	−0.574	0.043
Performance	0.564	0.137
Task 2		
Engagement	−0.323	−0.054
Performance	−0.497	−0.120
Task 3		
Engagement	0.034	−0.221
Performance	−0.401	−0.138
Task 4		
Engagement	0.456	0.121
Performance	0.671	0.038
Task 5		
Engagement	0.526	0.113
Performance	0.278	0.140

variance in Task 1 attention score, 74.1% and 13.7% of the variance in Task 2 attention score, 56% and 30.7% in Task 3, 75.9% and 13.4% in Task 4, and 80.3% and 9.1% in Task 5.

More specifically, the size of the dots in Appendix 3 represents the students' performance scores, whereas the bigger dots represent students who scored higher in the task. The red and blue dots each represent students' engagement and disengagement status (Goldhammer et al., 2016). The figures for Tasks 1, 4, and 5 showed clear alignments between the principal component scores and the problem-solving performance and engagement levels. For instance, visualization of the principal component scores for Tasks 1, 4, and 5 indicates a visible alignment between the size of the dot along the continuum of principal component score 1. Moreover, a clear alignment between the color coordination of the dots with principal component score 1, where the higher component score indicated an increased engagement level. However, the alignment between component scores and the performance and engagement level was less clear when visualized in Tasks 2 and 3, where the color coordination of the dots (engagement vs. disengagement) was less distinctive across the component scores.

The Pearson's correlation coefficients between the principal component scores and the examinee's performance and the engagement level also revealed similar findings. The primary component score in Task 4 and Task 5 showed moderate to high positive correlations coefficients with the students' engagement level (0.45–0.53) and the performance level (0.28–0.67). The primary component in Task 1, interestingly, showed moderate negative correlations with the engagement score (-0.57) and a positive correlation with the performance (0.564). We also observed that when the PCA scores aligned well with the engagement and the performance level, that comparably higher contribution to the prediction performance was observed. We discussed this and the implications of these findings further in the next section (Table 4).

Conclusions and discussion

The purpose of our study was to describe and demonstrate an analytic framework where the complex and long traces of process log data are used to understand the problem-solving skills and performance based on the examinee's log data in a problem-solving task in PIAAC 2012. Our engagement aware-LSTM model could outperform the other model in accurately classifying students based on their problem-solving performance.

The current empirical findings situate well in the existing literature by highlighting the importance of behavioral patterns or action sequences that are valuable to capture in modeling the examinee's problem-solving skills in PIAAC (He et al., 2019a, b). Some of the widely discussed benefits of incorporating behavioral patterns into problem-solving performance modeling involve the improvement of measurement accuracy (He et al., 2019a, b; Sireci & Zenisky, 2015), the establishment of the evidence to capture other latent or cognitive dimensions, such as engagement (He & von Davier, 2016; Zhu et al., 2016), and improvement in capturing abnormal behaviors (Hellas et al., 2017). Consistent with the previous literature, incorporating sequence-level process log features could successfully be associated with their performance (0.82–0.83 f1-score on average) while modeling students' engagement levels (0.92–0.97 f1-score on average) simultaneously in our findings. In our study, the low engagement that was captured across the problem-solving tasks could be interpreted as one source of anomalies that were commonly reported in the previous literature concerning formative or low-stakes assessments (Pastor et al., 2019; Pools & Monseur, 2021).

In addition, the findings from the current study align with previous research results indicating a close relationship between the examinee's engagement level and their problem-solving skills as well as the importance of modeling them together to have a better measure of students' problem-solving performance. Previously the connections between problem-solving performance and engagement were studied in relation to the complexity of the testing or assessment environments such as interactive games (Eseryel et al., 2014). For instance, Lein et al. (2016) indicated that engagement is a unique significant predictor that was associated with students' mathematical problem-solving performance when controlling for students' prior knowledge. Similarly, ongoing efforts are made in measurement research, where variations of IRT models are introduced to accurately estimate students' abilities (Nagy & Ulitzsch, 2022; Wise & DeMars, 2006).

Accordingly, the problem-solving task with the largest performance improvement in measuring students' problem-solving performance was in Task 1, Task 4, and Task 5, where the correlation coefficients between the performance and the engagement scores were the highest ($\rho = 0.480$, $\rho = 0.412$, $\rho = 0.373$). Conversely, in the tasks that showed a low to the negligible correlation between engagement and performance (2 and 3), the improvement in performance also remained relatively low.

Implication

The results provide practical and methodological implications for test developers and psychometric researchers. Using our approach, students' problem-solving abilities can be modeled in real-time and predicted to provide more direct and prompt feedback for

student performance. Also, the visualization and validation of the interim layer of complex machine learning models provide important evidence and insights to psychometric researchers which allows them to compare the model performance of deep learning models with the traditional psychometric approaches, such as IRT. Last, our engagement-aware model may allow test developers to adopt the system in a low-stakes assessment setting where the accurate evaluation of the student's ability, knowledge, and skills is challenging due to the lack of student motivation or engagement.

Wise and Kong, (2005) previously outlined large-scale assessment scenarios where the simultaneous measurement of engagement and students' ability level (e.g., problem-solving performance) may be recommended. First, the use of a low-stakes environment to pilot and validate the large-scale high-stakes exams may entail assessment situations where engagement detection may be necessary. Large-scale assessments, such as PIAAC and PISA commonly adopt such approaches to investigate the psychometric properties of the item prior to being officially introduced in their test booklets. Second, large-scale assessments are increasingly used to make inferences about teacher, school, and district evaluation, which may be deemed by the students to have low to negligible consequences for each participating individual. Not explicitly modeling students' engagement level during the participation may have significant consequences on validity of the test scores.

In essence, the deep learning methods proposed in this study provide the benefits of a data-informed and machine-learning based approach with an educational and psychometric consideration which could increase the capacity of promptly and accurately deriving decisions about examinee's performance from the education assessment with an increasingly digitized environment.

Limitations and future research

While our study was carefully constructed and implemented to avoid potential bias, we acknowledge that it is not free from limitations, which can be addressed in future research. First, the use of Principal Component Analysis (PCA) to improve the validity and interpretability of our model provided important benefits. However, it is important to recognize the limitations of PCA, notably its linear nature, which might not capture all non-linear relationships in the data. Also, the interpretation of the principal components, being linear combinations of original features, might not always be straightforward. Despite these limitations, the application of PCA on the last attention layer remains a valuable tool, offering a balanced approach to understanding and interpreting complex models in the context of educational assessments. Hence, we encourage future studies focusing on validating the PCA results to evaluate whether such patterns and relationships can be replicated and revealed when analyzing similar types of process data in large-scale assessment settings.

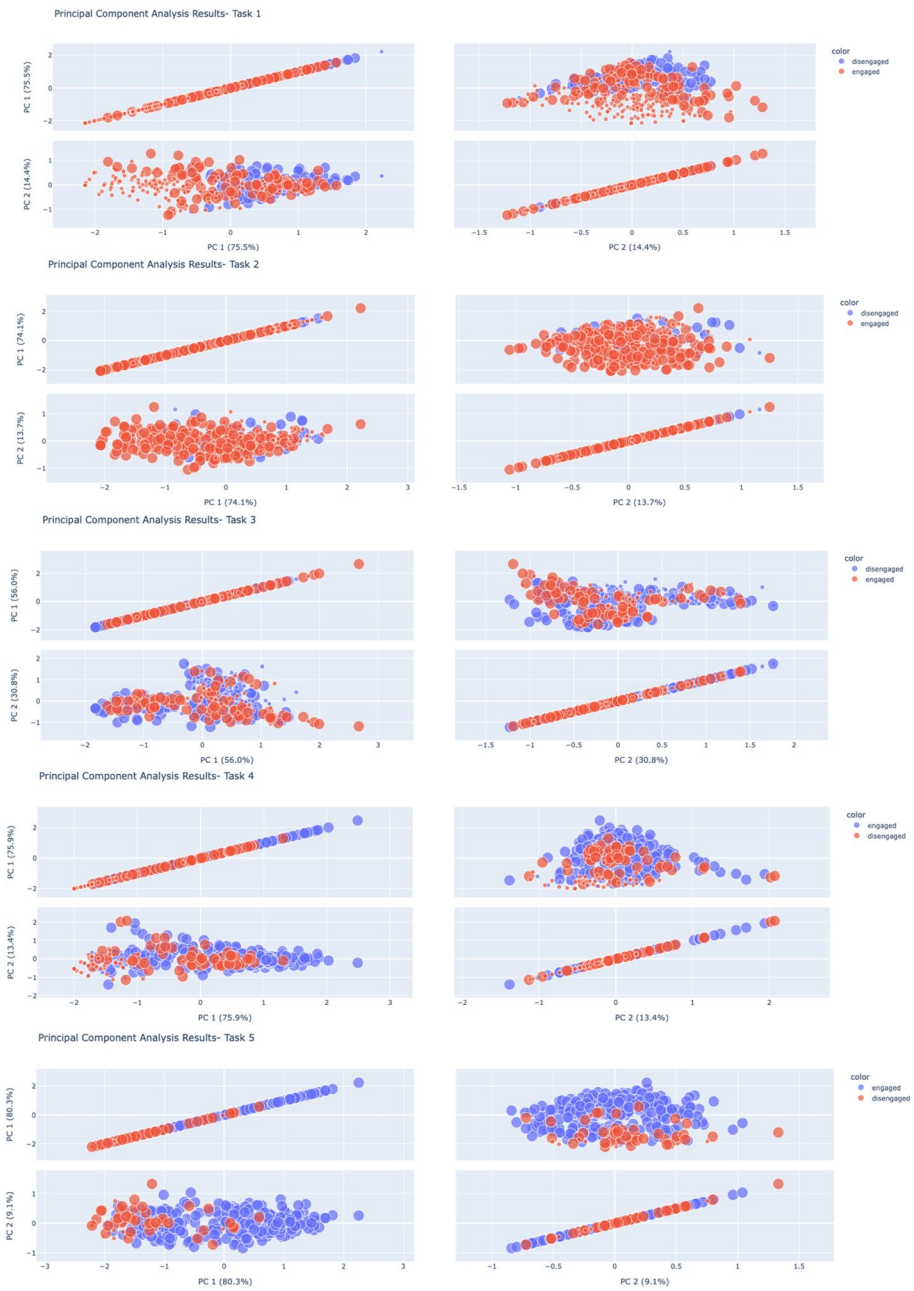
Appendix 1

Domains of Three Core Conceptual Dimensions. Adapted from "PIAAC Problem Solving in Technology-Rich Environments: A Conceptual Framework", OECD Education Working Papers, No. 36, OECD Publishing, Paris.

Dimension	Domain	Examples
Task	Purpose/context	Personal, Work/occupation, Civic purposes
	Intrinsic complexity	Minimal number of steps required to solve the problem
		Number of options or alternatives at various stages in the problem space
		Diversity of operators required, complexity of computation/transformation
		Likelihood of impasses or unexpected outcomes
		Amount of transformation required to communicate a solution
	Explicitness of problem statement	Ill-defined (implicit, unspecified) vs. well-defined (explicit, described in detail)
Technology	Hardware devices	Desktop or laptop computers, mobile phones, personal assistants, geographical information systems, integrated digital devices
	Software applications	File management, Web browser, Email, Spreadsheet
	Commands, functions	Buttons, Links, Textboxes, Copy/Cut-Paste, Sort, Find
	Representations	Texts, Sound, Numbers, Graphics (fixed or animated), Video
Cognitive dimension	Goal setting and progress monitoring	Identifying one's needs or purposes, given the explicit and implicit constraints of a situation
		Establishing and applying criteria for constraint satisfaction and achievement of a solution
		Monitoring progress
		Detecting and interpreting unexpected events, impasses and breakdowns
	Planning, self-organizing	Setting up adequate plans, procedures, and strategies (operators) and selecting appropriate devices, tools or categories of information
	Acquiring and evaluating information	Orienting and focusing one's attention; selecting information; assessing reliability, relevance, adequacy, comprehensibility; and reasoning about sources and contents
	Making use of information	Organizing information, integrating across potentially inconsistent texts and across formats, making informed decisions
		Transforming information through writing, from text to table, from table to graph, etc., and communicating with relevant parties

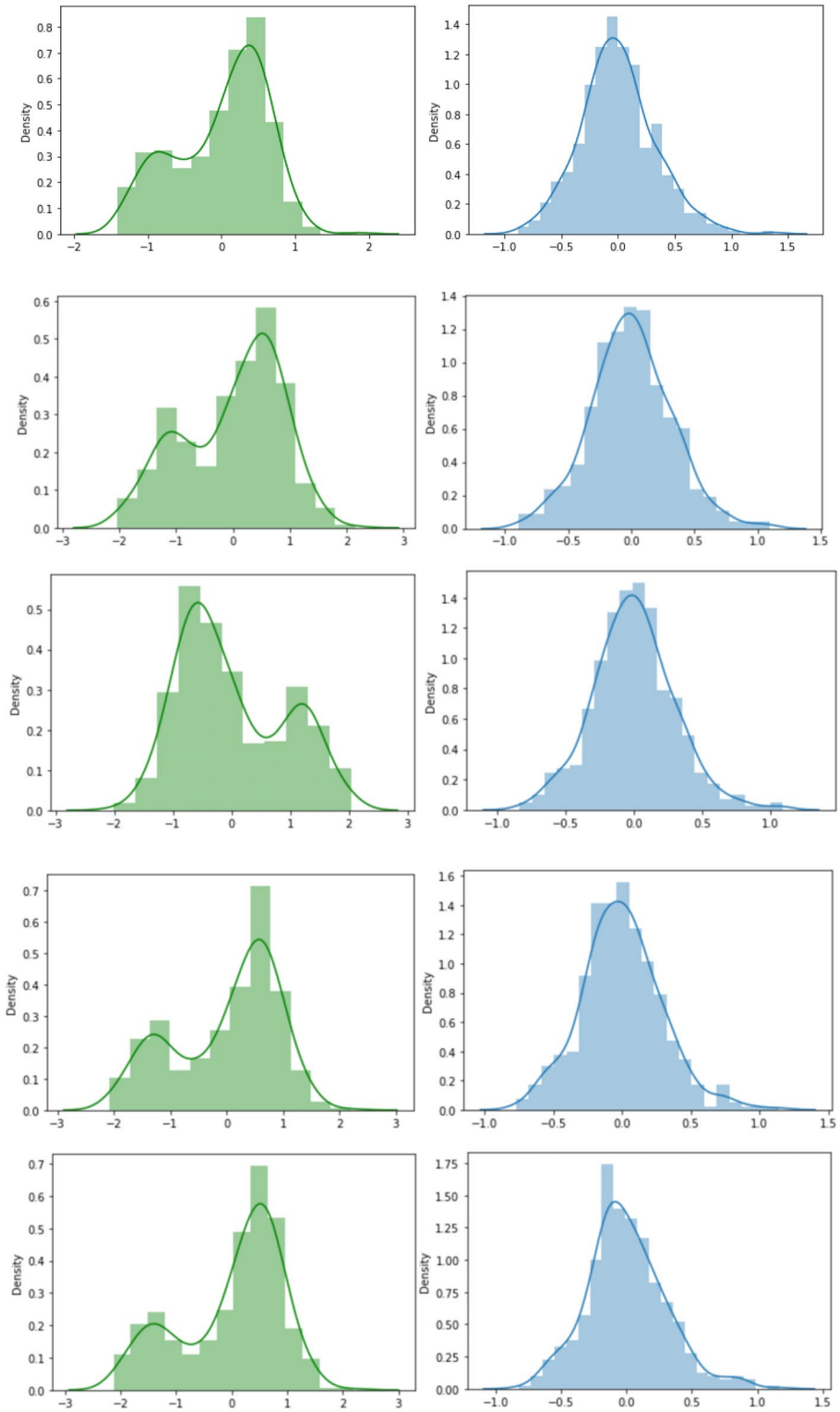
Appendix 2

Principal Component Analysis Results for the Task-level Attention Scores.



Appendix 3

Distribution of the principal component scores extracted from the attention model for Task 1 to Task 5 (green = PC 1, blue = PC 2).



Acknowledgements
Not applicable.

Author contributions
Data cleaning: JS, WP; Manuscript preparation: JS, BW, WP, MG; Data analysis: JS, BW.

Funding
Not applicable.

Availability of data and materials
The PIAAC PS-TRE 2012 Log dataset is available at https://www.oecd.org/skills/piaac/.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
All the authors consented for publication.

Competing interests
Not applicable.

Received: 30 March 2023 Accepted: 22 February 2024
Published online: 01 March 2024

References

Ai, F., Chen, Y., Guo, Y., Zhao, Y., Wang, Z., Fu, G., & Wang, G. (2019). Concept-Aware Deep Knowledge Tracing and Exercise Recommendation in an Online Learning System. *International Educational Data Mining Society*.

Barber, W., King, S., & Buchanan, S. (2015). Problem based learning and authentic assessment in digital pedagogy: Embracing the role of collaborative communities. *Electronic Journal of E-Learning, 13*(2), 59–67.

Braun, H., Kirsch, I., & Yamamoto, K. (2011). An experimental study of the effects of monetary incentives on performance on the 12th-grade NAEP reading assessment. *Teachers College Record, 113*(11), 2309–2344.

Chen, T., & Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining (pp. 785– 794). https://doi.org/10.1145/2939672.2939785

Chen, H., Huang, Y., & Nakayama, H. (2018, December). Semantic aware attention-based deep object co-segmentation. In Asian Conference on Computer Vision (pp. 435–450). Springer, Cham.

Deribo, T., Kroehne, U., & Goldhammer, F. (2021). Model-based treatment of rapid guessing. *Journal of Educational Measurement, 58*(2), 281–303.

Eseryel, D., Law, V., Ifenthaler, D., Ge, X., & Miller, R. (2014). An investigation of the interrelationships between motivation, engagement, and complex problem solving in game-based learning. *Journal of Educational Technology & Society, 17*(1), 42–53.

Fei, M., & Yeung, D. Y. (2015, November). Temporal models for predicting student dropout in massive open online courses. In *2015 IEEE international conference on data mining workshop (ICDMW)* (pp. 256–263). IEEE.

Goldhammer, F., Martens, T., Christoph, G., & Lüdtke, O. (2016). Test-Taking Engagement in PIAAC. *OECD Education Working Papers*, No. 133. OECD Publishing, Paris, https://doi.org/10.1787/5jlzfl6fhxs2-en.

Goldhammer, F., Naumann, J., Stelter, A., Tóth, K., Rölke, H., & Klieme, E. (2014). The time on task effect in reading and problem solving is moderated by task difficulty and skill: insights from a computer-based large-scale assessment. *Journal of Educational Psychology, 106*(3), 608.

Hanga, K. M., Kovalchuk, Y., & Gaber, M. M. (2020). A graph-based approach to interpreting recurrent neural networks in process mining. *IEEE Access, 8*, 172923–172938.

He, Q., & von Davier, M. (2015). Identifying feature sequences from process data in problem-solving items with n-grams. In *Quantitative Psychology Research: The 79th Annual Meeting of the Psychometric Society, Madison, Wisconsin, 2014* (pp. 173–190). Springer International Publishing.

He, Q., & von Davier, M. (2016). Analyzing process data from problem-solving items with n-grams: Insights from a computer-based large-scale assessment. *In Handbook of research on technology tools for real-world skill development* (pp. 750–777). IGI Global.

He, Q., Liao, D., & Jiao, H. (2019). Clustering behavioral patterns using process data in piaac problem-solving items. In *Theoretical and practical advances in computer-based educational measurement* (pp. 189–212). Springer, Cham.

He, Q., Borgonovi, F., & Paccagnella, M. (2019). *Using process data to understand adults' problem-solving behaviour in the programme for the international assessment of adult competencies (PIAAC): Identifying generalised patterns across multiple tasks with sequence mining*. OECD Education Working Papers No. 205.

Hellas, A., Leinonen, J., & Ihantola, P. (2017). Plagiarism in take-home exams: help-seeking, collaboration, and systematic cheating. *In Proceedings of the 2017 ACM conference on innovation and technology in computer science education* (pp. 238–243).

Hernández-Blanco, A., Herrera-Flores, B., Tomás, D., & Navarro-Colorado, B. (2019). A systematic review of deep learning approaches to educational data mining. *Complexity*, *2019*.

He, Q., Liao, D., & Jiao, H. (2019). *Clustering behavioral patterns using process data in piaac problem-solving items. Theoretical and practical advances in computer-based educational measurement.* 189–212.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation, 9*(8), 1735–1780.

Jiang, B., Wu, S., Yin, C., & Zhang, H. (2020). Knowledge tracing within single programming practice using problem-solving process data. *IEEE Transactions on Learning Technologies, 13*(4), 822–832.

Jiang, Y., Gong, T., Saldivia, L. E., Cayton-Hodges, G., & Agard, C. (2021). Using process data to understand problem-solving strategies and processes for drag-and-drop items in a large-scale mathematics assessment. *Large-Scale Assessments in Education, 9*, 1–31.

Keslair, F. (2018). Interviewers, test-taking conditions and the quality of the PIAAC assessment. OECD *Education Working Papers*, No. 191. OECD Publishing.

Kroehne, U., & Goldhammer, F. (2018). How to conceptualize, represent, and analyze log data from technology-based assessments? A generic framework and an application to questionnaire items. *Behaviormetrika, 45*, 527–563.

Kuang, H., & Sahin, F. (2023). Comparison of disengagement levels and the impact of disengagement on item parameters between PISA 2015 and PISA 2018 in the United States. *Large-Scale Assessments in Education, 11*(1), 4.

Lein, A. E., Jitendra, A. K., Starosta, K. M., Dupuis, D. N., Hughes-Reid, C. L., & Star, J. R. (2016). Assessing the relation between seventh-grade students' engagement and mathematical problem solving performance. *Preventing School Failure: Alternative Education for Children and Youth, 60*(2), 117–123.

Liao, D., He, Q., & Jiao, H. (2019). Mapping background variables with sequential patterns in problem-solving environments: an investigation of United States adults' employment status in PIAAC. *Frontiers in Psychology, 10*, 646.

Liu, Q., Huang, Z., Yin, Y., Chen, E., Xiong, H., Su, Y., & Hu, G. (2019a). Ekt: Exercise-aware knowledge tracing for student performance prediction. *IEEE Transactions on Knowledge and Data Engineering, 33*(1), 100–115.

Liu, Y., Li, Z., Liu, H., & Luo, F. (2019b). Modeling test-taking non-effort in MIRT models. *Frontiers in Psychology, 10*, 145.

Lloyd, S. (1982). Least squares quantization in PCM. *IEEE Transactions on Information Theory, 28*(2), 129–137.

Lundgren, E., & Eklöf, H. (2020). Within-item response processes as indicators of test-taking effort and motivation. *Educational Research and Evaluation, 26*, 275–301.

Mullis, I. V., Martin, M. O., Fishbein, B., Foy, P., & Moncaleano, S. (2021). Findings from the TIMSS 2019 problem solving and inquiry tasks. *Retrieved from Boston College, TIMSS & PIRLS International Study Center. website:* https://timssandpirls.bc.edu/timss2019/psi.

Nagy, G., & Ulitzsch, E. (2022). A multilevel mixture IRT framework for modeling response times as predictors or indicators of response engagement in IRT models. *Educational and Psychological Measurement, 82*(5), 845–879.

Organisation for Economic Co-operation and Development (OECD). (2012). *Literacy, numeracy and problem solving in technology-rich environments: Framework for the OECD survey of adult skills*. OECD Publishing.

Organisation for Economic Co-operation and Development (OECD). (2019). *Beyond proficiency: Using log files to understand respondent behaviour in the survey of adult skills*. OECD Publishing. https://doi.org/10.1787/0b1414ed-en

Pastor, D. A., Ong, T. Q., & Strickman, S. N. (2019). Patterns of solution behavior across items in low-stakes assessments. *Educational Assessment, 24*(3), 189–212.

Polyak, S. T., von Davier, A. A., & Peterschmidt, K. (2017). Computational psychometrics for the measurement of collaborative problem solving skills. *Frontiers in Psychology, 8*, 2029.

Pools, E., & Monseur, C. (2021). Student test-taking effort in low-stakes assessments: Evidence from the English version of the PISA 2015 science test. *Large-Scale Assessments in Education, 9*(1), 1–31.

Qiao, M., & Li, H. (2020, October). Application of PCA-LSTM model in human behavior recognition. In Journal of Physics: Conference Series (Vol. 1650, No. 3, p. 032161). IOP Publishing.

Ramalingam, D., & Adams, R. J. (2018). How can the use of data from computer-delivered assessments improve the measurement of twenty-first century skills? In E. Care, P. Griffin, & M. Wilson (Eds.), *Assessment and teaching of 21st century skills* (pp. 225–238). Springer International Publishing.

Rouet JF, Betrancourt M, Britt MA, Bromme R, Graesser AC, Kulikowich JM, Leu DJ, Ueno N, Van Oostendorp H. (2009). PIAAC Problem Solving in Technology-Rich Environments: A Conceptual Framework. OECD Education Working Papers, No. 36. *OECD Publishing (NJ1)*.

Rocktäschel, T., Grefenstette, E., Hermann, K. M., Kočiský, T., & Blunsom, P. (2015). Reasoning about entailment with neural attention.https://arxiv.org/abs/1509.06664

Sahin, F., & Colvin, K. F. (2020). Enhancing response time thresholds with response behaviors for detecting disengaged examinees. *Large-Scale Assessments in Education, 8*(1), 1–24.

Salles, F., Dos Santos, R., & Keskpaik, S. (2020). When didactics meet data science: process data analysis in large-scale mathematics assessment in France. *Large-Scale Assessments in Education, 8*(1), 1–20.

Scherer, R., Greiff, S., & Hautamäki, J. (2015). Exploring the relation between time on task and ability in complex problem solving. *Intelligence, 48*, 37–50.

Schnipke, D. L. (1996). *Assessing speededness in computer-based tests using item response times*. Baltimore: The Johns Hopkins University.

Shin, J., Chen, F., Lu, C., & Bulut, O. (2022). Analyzing students' performance in computerized formative assessments to optimize teachers' test administration decisions using deep learning frameworks. *Journal of Computers in Education, 9*(1), 71–91.

Sireci, S. G., & Zenisky, A. L. (2015). Computerized innovative item formats: Achievement and credentialing. *In Handbook of test development* (pp. 329–350). Routledge.

Tang, S., Peterson, J. C., & Pardos, Z. A. (2016, April). Deep neural networks and how they apply to sequential education data. In *Proceedings of the third (2016) acm conference on learning@ scale* (pp. 321–324).

Organisation for Economic Co-operation and Development (OECD). (2016). *Technical report of the* survey *of adult skills (PIAAC)*. 2nd Edition.

Terrin, N., Schmid, C. H., Griffith, J. L., D'Agostino, R. B., Sr., & Selker, H. P. (2003). External validity of predictive models: a comparison of logistic regression, classification trees, and neural networks. *Journal of Clinical Epidemiology, 56*(8), 721–729.

Ulitzsch, E., He, Q., Ulitzsch, V., Molter, H., Nichterlein, A., Niedermeier, R., & Pohl, S. (2021). Combining clickstream analyses and graph-modeled data clustering for identifying common response processes. *Psychometrika, 86*(1), 190–214.
Ulitzsch, E., Ulitzsch, V., He, Q., & Lüdtke, O. (2022). A machine learning-based procedure for leveraging clickstream data to investigate early predictability of failure on interactive tasks. *Behavior Research Methods, 55*, 1–21.
Van Laar, E., Van Deursen, A. J., Van Dijk, J. A., & De Haan, J. (2017). The relation between 21st-century skills and digital skills: a systematic literature review. *Computers in Human Behavior, 72*, 577–588.
Von Davier, A. A. (2017). Computational psychometrics in support of collaborative educational assessments. *Journal of Educational Measurement, 54*(1), 3–11.
Vanek, J. (2017). Using the PIAAC framework for problem solving in technology-rich environments to guide instruction: An introduction for adult educators. Washington: PIAAC
Wang, L., Sy, A., Liu, L., & Piech, C. (2017, April). Deep knowledge tracing on programming exercises. In *Proceedings of the fourth (2017) ACM conference on learning@ scale* (pp. 201–204).
Wang, K. D., Salehi, S., Arseneault, M., Nair, K., & Wieman, C. (2021, June). Automating the Assessment of Problem-solving Practices Using Log Data and Data Mining Techniques. In *Proceedings of the Eighth ACM Conference on Learning@ Scale* (pp. 69–76).
Wise, S. L. (2015). Effort analysis: Individual score validation of achievement test data. *Applied Measurement in Education, 28*(3), 237–252.
Wise, S. L. (2017). Rapid-guessing behavior: Its identification, interpretation, and implications. *Educational Measurement: Issues and Practice, 36*(4), 52–61.
Wise, S. L. (2020). Six insights regarding test-taking disengagement. *Educational Research and Evaluation, 26*(5–6), 328–338.
Wise, S. L., & DeMars, C. E. (2006). An application of item response time: The effort-moderated IRT model. *Journal of Educational Measurement, 43*(1), 19–38.
Wise, S. L., & Kong, X. (2005). Response time effort: A new measure of examinee motivation in computer-based tests. *Applied Measurement in Education, 18*(2), 163–183.
Zeng, W., Li, J., Quan, Z., & Lu, X. (2021). A deep graph-embedded LSTM neural network approach for airport delay prediction. *Journal of Advanced Transportation, 2021*, 1–15.
Zhang, T., & Guo, G. (2020). Graph attention LSTM: A spatiotemporal approach for traffic flow forecasting. *IEEE Intelligent Transportation Systems Magazine, 14*(2), 190–196.
Zhang, Z., Lv, Z., Gan, C., & Zhu, Q. (2020). Human action recognition using convolutional LSTM and fully-connected LSTM with different attentions. *Neurocomputing, 410*, 304–316.
Zhu, M., Shu, Z., & von Davier, A. A. (2016). Using networks to visualize and analyze process data for educational assessment. *Journal of Educational Measurement, 53*(2), 190–211.

Publisher's Note