


How Difficult is the Task for you? Modelling and Analysis of Students' Task Difficulty Sequences in a Simulation-Based POE Environment

Sadia Nawaz¹ , Namrata Srivastava¹, Ji Hyun Yu², Anam Ahmad Khan¹
Gregor Kennedy¹, James Bailey¹ and Ryan Shaun Baker³

¹ The University of Melbourne, Parkville, VIC 3010, Australia,
nawazs@student.unimelb.edu.au

² The University of Michigan, Ann Arbor, MI 48109, United States

³ University of Pennsylvania, PA 19104, United States

Abstract. Task difficulty (TD) reflects students' subjective judgement on the complexity of a task. We examine the TDs data of 236 undergraduate students in a simulation-based *Predict-Observe-Explain (POE)* environment using three different labels *easy*, *medium* and *hard*. Generally, the students who perceive the tasks to be *easy* or *hard* perform poorly at the transfer task than the students who perceive the tasks to be *medium* or moderately difficult. Sequences of students' TDs are analysed which consist of a set of several judgements, collected once for each task in a *POE* sequence. The analysis suggests that given a sequence of TDs, difficulty level *hard* followed by a *hard* may lead to poorer learning outcomes at the transfer task. By contrast, difficulty level *medium* followed by a *medium* may lead to better learning outcomes at the transfer task.

In terms of the TD models, we identify student behaviours that can be reflective of their perceived difficulties. Generally, the students who report that the tasks are *easy*, adopt a trial-and-error behaviour where they spend lesser time and make more attempts on tasks. By comparison, the students who complete the tasks in a longer time by making more attempts are likely to report that the following task is *hard*. For the students who report *medium* TDs, mostly these students seem to reflect on tasks where they spend a long time and require fewer attempts for task completions. Additionally, these students provide longer texts for explaining their hypothesis reasoning.

Understanding how student behaviours and TDs manifest over time and how they impact students' learning outcomes is useful, especially when designing for real-time educational interventions, where the difficulty of the tasks could be optimised for students. It can also help in designing and sequencing the tasks for the development of effective teaching strategies that can maximise students' learning.

Keywords: Task difficulty, Task complexity, Predict-Observe-Explain, Learning outcomes, *L*-statistic, Bi-gram sequences, Modelling, Intervention, Flow, Zone of Proximal Development.

1 Introduction

Students' perceptions of tasks or their task difficulties (TDs) can influence their learning behaviours [5, 8]. For example, when a task is challenging yet attainable, students may invest effort and persist at it. In contrast, students may not engage in a task if they repeatedly fail at it [35, 71]. This, then, engenders the question: how can instructors design for optimal learning conditions where students get challenged but feel confident in accomplishing the tasks? To address this question, we analyse the relation of students' task difficulties with their learning outcomes (e.g., is it more probable for the *high* achievers to report that the TDs are *easy* or is it the other way around). Further, we observe how TDs vary in a simulation-based learning environment (e.g., how likely it is for TDs to transition from *easy* to *hard* and vice-versa). Then, we assess whether students' sequences of TDs can be indicative of their learning outcomes (i.e., we examine students' TD sequences to identify which sequences might be better in terms of students' achievements). Lastly, we build and analyse the detectors or models of students' TDs to identify if there are certain task-based interaction patterns (such as students' time on tasks, task attempts, length of students' textual responses and the nature of students' prior knowledge) that may be associated with students' perceptions of difficulties or their TDs.

In this paper, TDs are analysed in a digital simulation-based *Predict-Observe-Explain (POE)* learning environment by using the likelihood statistic (*L*-stat). The AIED community has frequently used *L*-stat for studying students' affective dynamics [25, 26, 28, 29, 49, 50]. Compared to a traditional classroom environment, a benefit of analysing TDs in a digital setting is that students can receive just-in-time support. For instance, task complexity can be adjusted by the instructors to match students' level of understanding or individual students may also choose and change the level of TDs in a self-controlled setting [3, 32, 43, 88]. A better understanding of students' TDs can enable interventions that can improve students' learning [1, 76, 79] and reduce undesirable behaviours such as gaming the system [2] and task disengagement [39].

2 Related Work

Task complexity and task difficulty are often used interchangeably. However, they are two different constructs [74, 75]. Task complexity represents the characteristics or cognitive demands of a task [14]. A task which requires more cognitive resources is a complex task, whereas a task which requires lesser cognitive resources is considered a simple task. By comparison, task difficulty refers to the task-doers' perceived difficulty or their subjective judgment in terms of the effort which is needed to complete the tasks. In this paper, we use perceived difficulty and task difficulty (TD) interchangeably.

Different learners can perceive the same tasks differently [14]. Researchers have shown that TDs can influence students' motivation [45] and self-regulation [5]. TDs can also affect students' problem-solving strategies and tactics. For example, DeLoache, Cassidy and Brown [31] suggest that "problems that are too *easy* or too difficult are less likely to elicit strategic behaviour than the problems that present a

moderate degree of challenge" (1985, p. 125). Further, the "law of optimum perceived difficulty" states that, if the tasks are perceived very *easy* or very *hard*, they can result in lower levels of engagement than the moderately difficult tasks – which may lead to higher levels of engagement [8]. Vygotsky [85] suggested that for instruction to be effective it must be aimed at learners' proximal level of development (where learners can succeed with assistance; a difficulty that is somewhat more challenging than an exact match to a student's skill level, but not so challenging that the student cannot succeed). Csikszentmihalyi, in his works [21, 82] talks about TDs and their influence on emotions. He suggests that a person may feel worried and anxious when presented with overly challenging tasks and may feel bored if the tasks are too *easy*. However, when the tasks are moderately difficult, or they offer just the right challenge, a positive 'flow' experience may occur [22, 23]. Therefore, different emotions can be encountered based on how an individual perceives a given task.

This, then raises the question: what relation do TDs have to students' learning outcomes? The data is not entirely clear on these theoretical perspectives. Some studies report that TDs have a negative association with students' self-efficacy and performance [60, 62], yet [10] states that 'certain difficulties can enhance learning'. There have been a number of studies indicating that students can learn from challenges that lead them to identify and articulate their current views, examine their ideas and clarify their misconceptions [47, 48]. To sum up, in this paper, we investigate the following questions:

RQ1: What relation do task difficulties have with students' learning outcomes?

RQ2: How do task difficulties vary over time?

RQ3: Is there a sequence of task difficulties that is indicative of better learning?

RQ4: Are there any interaction patterns that are important in determining students' levels of task difficulties?

3 Learning Environment

3.1 Predict Observe Explain (POE) Simulations

Students' prior knowledge is often based on their daily life observations, which may differ from the most precise scientific conceptions. The difference between students' prior conceptions and the true scientific conceptions is called alternative conceptions or misconceptions [84]. Researchers in science education generally agree that to assist students in learning new scientific knowledge; they should be made aware of their prior knowledge [40, 47, 48]. "Diagnostic" teaching strategies should be applied to confront students' misconceptions [7]. One such framework that considers students' prior conceptions is the *Predict-Observe-Explain (POE)* instructional design [86]. *POE* is a three-phase, iterative design [30].

1. During the *Prediction* task, students formulate a hypothesis. They are often asked to provide the reasons as to why they committed to it.
2. During the *Observation* task, students can test their hypotheses by changing parameters or variables in a simulation. They can see the effects of their manipulations.

This phase is especially crucial for the students who propose incorrect hypotheses, as they can then see a mismatch between what they predicted and what they are observing [33].

3. During the *Explanation* phase, clarifications are provided to students detailing the relationship between variables or parameters representing the conceptual phenomenon under investigation. This phase assists students to reconcile any discrepancies between their predictions and observation in the simulation [44].

POEs can be applied in face-to-face, online and computer lab contexts [20]. They can promote student discussion [86], probe into their prior knowledge and help them update their prior conceptions [19, 53, 83]. *POE* learning designs can make digital environments more engaging by offering autonomy to students and allowing them to complete the tasks at their own pace [53, 81]. Rather than showing the solution to students, *POEs* encourage the students to solve the problems for themselves by engaging in tasks that are high in cognitive demands. As students engage in demanding tasks, they can undergo various emotions. While recently, there have been *POE*-based studies that analyse students' affective experiences, their struggle and confusion [52, 64, 65]. There is a need to understand students' task difficulties, especially when students reach an impasse, failure or when they face challenges.

To the best of our knowledge, TDs have not yet been investigated within *POE* based simulation environments. Understanding how students' TDs manifest over time, how they relate to students' task-based interaction patterns and how they can impact students' learning outcomes is useful, especially when designing for real-time educational interventions. Therefore, it is essential that we examine how TDs vary in these environments.

3.2 Course and Module Description

The data in this study is taken from an online project-based **course** called *Habitable Worlds*. It aims to introduce the foundational concepts of Physics, Chemistry and Biology [46]. It intends to develop problem-solving and logical reasoning skills in students through immersive and interactive tasks in a guided discovery environment. *Habitable Worlds* is built using Smart Sparrow's eLearning platform¹, which records moment by moment activity of students. The learning environment in this program is 'adaptive'. It allows the provision of feedback and hints based on students' responses (or lack of responses). It also typically means that the students are not allowed to progress or move on until a task has successfully been completed. Furthermore, for students who seem to hold misconceptions, there is occasional pathways adaptivity where students are taken to additional tasks or screens to provide them with extra material that can support in rectifying their prior conceptions.

Habitable Worlds is offered to undergraduate students over a duration of 7.5 weeks, and it consists of 67 interactive **modules**. The current study focuses on an introductory module called *Stellar Lifecycles*. The concept under investigation is the relation between a star's mass and its lifespan. There are several **tasks** within this module which

¹ <https://www.smartsparrow.com/research/>

involve one or more of the following activities: providing free-text answers to a question, watching videos, responding to multiple-choice questions or the 'submissions' associated with simulations. In this module, students follow the prescribed sequence of tasks or activities. However, as discussed above, there is occasional pathways adaptivity offered for the remediation of students who make errors.

3.3 Tasks Description

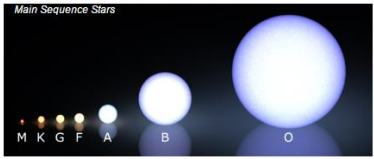
Of the 23 tasks within this module, we utilise the following *POE* based tasks:

- *Prediction*: Students need to select a hypothesis from five possible choices regarding the relationship between stellar mass and stellar lifespan (see *Figure 1*). Then, they need to report their reasons (through free text responses) for selecting that hypothesis.

Stellar Lifecycles

How do you think mass and stellar lifetime are related for Main Sequence stars?

Main Sequence Stars



There is no relationship between mass and lifetime of a star.

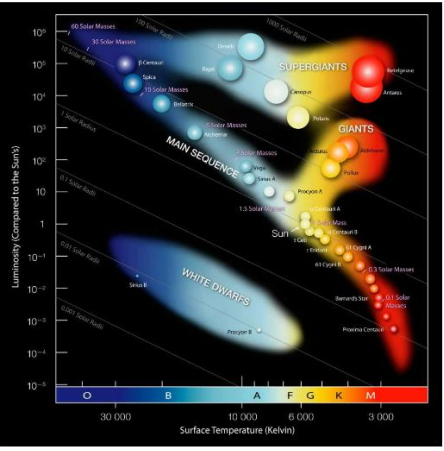
Low-mass stars live longer than high-mass stars.

Medium-mass stars live longer than either low-mass and high-mass stars.

Medium-mass stars live shorter than either low-mass and high-mass stars.

High-mass stars live longer than low-mass stars.

Why do you think this is the case?



TEST HYPOTHESIS

Figure 1: *Prediction* task – students are asked to propose a hypothesis and provide reasoning for their choice of hypothesis.

- *Observation 1*: During the first stage of the *Observe* task, students explore the stellar nursery simulator to create virtual stars, manipulate their mass and run them (as many times as they wish). Through this simulator, students can study and hopefully understand the relation between stellar mass and its lifespan.
- *Observation 2*: During the second stage of the *Observe* task, students need to create at least three different stars within a specified mass range. Then, they need to record

the mass and associated lifespan of these stars. Next, given their observations, they need to either accept or reject their earlier proposed hypotheses.

- *Explanation 1*: This task is only available to those students who make incorrect predictions and endorse them or those who make correct predictions but reject them. This task can assist students in rectifying their prior hypotheses.
- *Explanation 2*: This task requires the students to report the minimum and the maximum lifespan of seven different stellar classes. Students can again create and run stars within the stellar nursery simulator. Most students seem to struggle at this task as they need to manipulate several different stellar classes. This struggle is reflected in students' making repeated attempts. Those who manipulate only one stellar class at a time (more systematic) are more likely to complete this task than those who manipulate more than one stellar classes (less systematic) [65].
- *Post POE*: At the final stage of the POE sequence, students are provided with a short lecture-style video to explain to them why low mass stars live longer and how a star's mass and internal pressure contribute to the nuclear fusion process which fuels the burning of stars and hence their lifespan.
- After the *POE* sequence of tasks, students make observations of different stars as they burn. They are then asked to answer multiple-choice questions and report on the changes in the stellar classification of burning stars.
- *Transfer Task*: After completing the *Stellar Lifecycles* module, students need to complete the knowledge-transfer task – the *Stellar Applications* module. At this task, students are asked to calculate the properties (such as mass, luminosity, lifespan, and temperature) of six stars. Students also need to identify the longest- and shortest-lived stars (see *Figure 2*). While students can make multiple attempts at this task, they are penalised by two marks for each incorrect attempt.

Unlike the *Stellar Lifecycles* module which consists of 23 different tasks, the *Stellar Applications* module consists of a single task. Therefore, for *Stellar Applications*, there is no *POE* sequence to be followed by the students. For calculating various stellar properties, students are only required to apply the formulae that were already introduced to them.

3.4 Participants

The data in this study is taken from the October 2017 offering of the course *Habitable Worlds*. A total of 236 non-science major undergraduate students attempted this module. Of these students, 50% were females, 46% were males, and 4% did not respond. In terms of age, 33% of students were younger than 20, 46% were between the age range of 21 and 30 both inclusive. The remaining 21% were older than 30.

In terms of ethnicity, 69% were 'White', 17% were Hispanic/Latino, and the remaining 14% were classified as 'Others' which included 'Asian', 'African American', 'American Indian/Alaska Native', 'non-resident alien' and 'two or more races'. The reason for combining these ethnicities was their small sample size. However, it should be noted that the 'Others' group is too diverse to be able to draw useful conclusions and therefore

in the results the focus should be on the groups that are large enough to be considered separately.

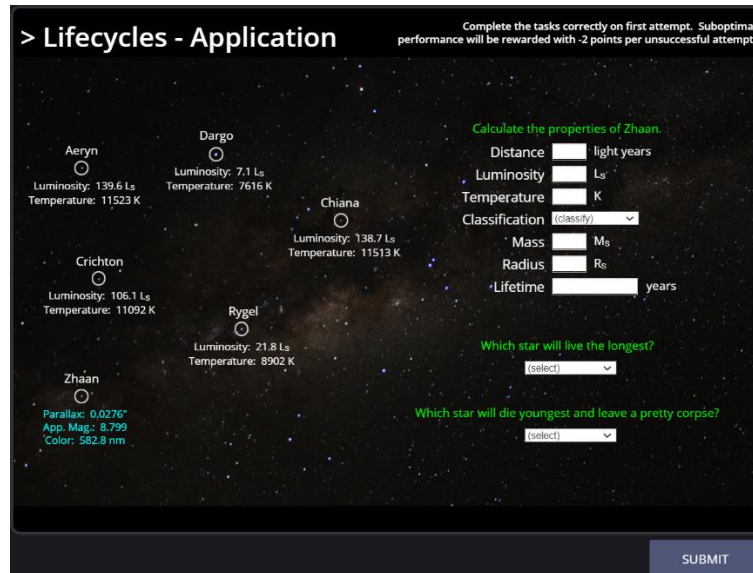


Figure 2: *Stellar Applications* module – to test the students based on the previously introduced concepts.

3.5 Measures

Correctness of hypotheses. To infer students' prior knowledge, we used their choice of hypothesis at the *Prediction* task. During this task, students are asked to make a prediction about a conceptual phenomenon relating the relationship between a star's mass and its lifespan. Out of five available options, only one hypothesis is correct, i.e., the low mass stars live longer than the high mass stars (see *Figure 1*).

Learning Outcomes. We analyse students' scores at the transfer task – the *Stellar Applications* module, which immediately follows the *Stellar Lifecycles* module. It tests students on the concepts that were already introduced to them. The maximum achievable score at this task is ten (10), and with each incorrect attempt, students are penalised by two (2) marks.

Perceived difficulty during-task. During each phase of the *POE* tasks, to infer students' perceived difficulty, they are asked to report their levels of confidence and challenge on a 6-point Likert scale: from 1 (not at all) to 6 (extremely). The following questions are asked:

- How confident are you that you understand the task right now?
- How challenging do you find the task right now?

Perceived difficulty after-task. At the end of the *POE* cycle, students can report their overall confidence and overall challenge on a 6-point scale. This is done only one time at the end of the *POE* sequence. Students are asked these questions:

- Overall, how confident are you that you understood the material in the preceding tasks?
- Overall, how challenging was the material in the preceding tasks?

The response to these survey items is voluntary. In terms of participation, *during-task*, 186 students report their perceived difficulty during the *Prediction* task, 151 and 146 students report their TDs during the *Observe-1* and *Observe-2* tasks respectively, 74 and 146 students report during the *Explain-1* and *Explain-2* tasks. Lastly, 185 students report their perceived difficulty *after-task*.

Trace data. In this naturalistic study, in addition to students' perceived difficulties (from above), we also utilised their trace data at a given task to model for their TDs at the subsequent tasks. Trace data or log-files are created when students take actions within the digital or online learning environment, e.g., when they open a module, when they make a submission associated with tasks or questions, or when they finish a learning session by signing out of the system [12, 42].

4 Data Pre-processing

4.1 Levels of Task Difficulty

For analysing students' task difficulties, we include those students who respond to one or more of the task-based surveys. As mentioned, survey items are related to students' confidence and challenge for a given task. To infer TDs, we assign the following three (3) labels:

- *Easy (E)*: if reported confidence exceeds reported challenge,
- *Hard (H)*: if reported confidence is lower than the reported challenge,
- *Medium (M)*: if reported confidence matches with the reported challenge

Note that our TD labels match with Csikszentmihalyi's flow theory [24]. While the flow theory frames students' affective experiences in terms of their challenge and skills, we use these measures (challenge and confidence) to infer students' perceptions of difficulties or TDs.

4.2 Task Difficulties and Students' Demographics

We started the analyses by investigating the association between students' demographics and their reported TDs. Knowledge of the underlying population is important as it can allow educators another opportunity to know their students and to monitor their progress and engagement. Several recent papers have raised the concern that research in AIED and other communities may be creating inequities by ignoring

important demographic differences between communities, and have called for publishing these types of analyses to verify whether there are group differences [4, 51, 61, 66, 67].

In this regard, we perform comparisons between genders (i.e., we compare the perceived difficulty of male and female students). We also compare students from different ethnic backgrounds. Due to small sample sizes, the students who declared themselves as 'Asian', 'African American', 'American Indian/Alaska Native', 'non-resident alien' and 'two or more races' are reported under the category 'Others'. Overall, to compare students' ethnicity and their TDs, we perform comparisons between 'White', 'Hispanic/Latino' and 'Others'. However, the 'Others' group is too diverse to be able to draw useful conclusions and therefore, in our results the focus should mostly be on the groups who are large enough to be considered separately.

Lastly, we compare students from different age groups (e.g., students between the age of 17-20, between 21-30 and lastly, the students above the age of 30). Comparisons are made at each of the *POE* tasks and separately for each level of TDs, using the Pearson's Chi-Square test (or the Fisher's exact test when the entries in the contingency table are less than five). Only significant or marginally findings are reported here.

4.3 Task Difficulties and Learning Outcomes

Learning outcomes are students' scores at the knowledge-transfer task. As described above, the transfer task immediately follows the *Stellar Lifecycles* module, and it tests students on the concepts that were already introduced to them. The maximum achievable score on the transfer task is ten (10), and for each repeated attempt two (2) points are deducted. *High* achieving students are those who scored above the mean ($M=9.21$, $SD=0.92$), while the students scoring below the mean are considered *low* achievers ($M=3.64$, $SD=4.58$).

To compare the above two student groups, we perform Pearson's Chi-square test (or Fisher's exact test when the entries in the contingency table are less than 5). Comparisons are presented for each level of TD and during each phase of the *POE* cycle.

4.4 Correctness of Hypotheses and Learning Outcomes

As mentioned, the *Stellar Lifecycles* module is aimed to introduce the students of the relationship between stellar mass and stellar lifespan. Presumably, students are not introduced to this concept prior to this module. We anticipated that many students might hold a misconception about this relation. Therefore, based on students' choice of hypothesis during the *Prediction* task, we infer students' levels of prior knowledge; and then, we compare the *high* and the *low* achieving students in terms of their prior knowledge.

4.5 Task Difficulty Transitions

During each phase of the *POE* tasks, as students report their confidence and challenge, we infer their TDs. Later, we use these TDs to estimate the likelihood statistics (*L*-stat) as well as the TD bi-gram sequences.

Calculating *L*-stat. After obtaining students' TDs, we compute the likelihood of transitions between any two possible states using the transition metric *L* [28], with self-transitions included in the calculation. This metric specifies the probability of a transition from a level at time *t* to *t*+1, after correcting for the base rate at time *t*+1. We can represent this as $L(\text{difficulty}_t \rightarrow \text{difficulty}_{t+1})$, where difficulty_t is the difficulty level at the current task and difficulty_{t+1} is the difficulty level at the next task:

$$L(\text{difficulty}_t \rightarrow \text{difficulty}_{t+1}) = \frac{P(\text{difficulty}_{t+1} | \text{difficulty}_t) - P(\text{difficulty}_{t+1})}{1 - P(\text{difficulty}_{t+1})}$$

To simplify, for difficulty levels *A* and *B*, the transition likelihood from $A \rightarrow B$ is:

$$L(A \rightarrow B) = \frac{P(B|A) - P(B)}{1 - P(B)}$$

Where $P(B)$ is the probability that difficulty level *B* occurs as a next state. Here, the first occurrence of any perceived difficulty is excluded from the calculation, as this occurrence cannot be considered for the next state. The conditional probability $P(B|A)$ is:

$$P(B|A) = \frac{\text{count}(A \rightarrow B)}{\text{count}(A)}$$

Here, $\text{count}(A \rightarrow B)$ is the number of times a difficulty level transitions from *A* to *B*, and $\text{count}(A)$ is the number of times the difficulty level *A* occurs as a previous state. The value of *L* may vary from $-\infty$ to 1. For a given transition, $A \rightarrow B$, if $L \approx 0$, we say that the transition occurs at chance level, if $L > 0$, we say that state *B* follows state *A*, above chance. Finally, if $L < 0$ then state *B* follows state *A* below chance [27].

For calculations, the *L*-statistic is computed separately for each student and for each possible transition. The transitions where *L* is undefined are excluded from further analysis. Later, one-sample (two-tailed) t-tests are conducted on the calculated *L* values to measure whether each transition is significantly more or less likely than chance. Next, the Benjamini-Hochberg (BH) post-hoc correction is applied to control for false positives, as the analysis involves multiple comparisons [49].

Generating TD bi-gram sequences. Here we analyse students' TDs as bi-grams, i.e., sequences of two consecutive tasks. Like bi-grams TDs can be analysed in the sequence of three consecutive tasks, four consecutive tasks or all tasks. However, as the length of the TD sequence increased, the sample size reduced correspondingly and hence it was decided to only consider the bi-gram sequences.

For bi-gram analysis, we considered only those students who responded to all task-based surveys and who also attempted the knowledge-transfer task – there were 63 such students. In this regard, given a sequence: '*easy-medium-medium-hard-hard-easy*', the associated bigrams are: '*easy-medium*', '*medium-medium*', '*medium-hard*', '*hard-hard*' and '*hard-easy*'. After this, we compare the students who report a given bigram sequence versus those who do NOT report it. For this, we perform t-tests and report the results in terms of *p-value* statistic and *t-value* statistic. Test result is considered significant if *p-value* < 0.05 (*) and marginally significant if *p-value* < 0.10 (•). As this analysis also involves multiple comparisons, BH post-hoc correction is applied.

4.6 Task Difficulty Modelling and Feature Extraction

To understand what interaction patterns are important in determining students' level of difficulties, we develop models. These models or detectors are developed to predict students' TDs using their interaction data as well as their perceived difficulties from the preceding tasks. For feature extraction, we utilise students' trace data from the simulation-based learning environment. Overall, *Habitable Worlds* course had a total of 613,653 task-based interactions recorded in the system; of these, 10,422 interaction entries were related to *Stellar Lifecycles*.

Handling outliers. As a first step towards data pre-processing, outliers are eliminated. All those interaction entries where students' time on task exceeds 60 minutes are considered outliers. Eliminating such entries is important as it can be that students start a learning session and then leave the browser window open without meaningfully engaging in a learning activity.

Feature extraction. After removing the outliers, we extract the features using students' interaction logs at each of the POE tasks. These features represent student actions or activities, e.g., the number of errors that students make on a given task, the time that students take to complete a task, and the count of attempts made by students. Based on the learning design, there are some features that are task-related, e.g., during the Prediction task, we include information whether students' proposed hypotheses are correct or not, the length of textual reasoning entered by students, as well as the sentiment analysis of students' textual reasoning in terms of positive, negative, and neutral terms. In addition to the above interaction data, we also use students' perceptions of difficulties – TDs in the preceding tasks to predict their TDs at the subsequent tasks. The definition of the various features, used for TD modelling, is provided at the end of the paper, in appendix *Table A1*.

Classification models. We develop separate models for each level of TD, e.g., *hard* is distinguished from *not_hard* (which consists of TDs: *easy* and *medium*). Similarly, *easy* is distinguished from *not_easy* (consisting of TDs: *hard* and *medium*), and lastly, *medium* is distinguished from *not_medium*, (which consists of TDs: *easy* and *hard*). Models are developed during each task of the *POE* cycle. For modelling we encode the

TD predictor variables as binary, e.g., when the detectors for TD *easy* are considered, we analyse how reporting *easy* or *not_easy* on a current task can affect reporting *easy* on a subsequent task. Similarly, when the detectors for TD *medium* are considered, we analyse how reporting *medium* or *not_medium* on a current task can affect reporting *medium* on a subsequent task. For *Prediction* task, which is the first in a *POE* sequence, data from a prior task is used. This task asked students to calculate the mass and the radius of different stars. On this task, for calculating mass and radius, students can access the required formulae through clickable hints.

To develop TD models or detectors, we use logistic regression [9]. Feature distillation is conducted using the backward stepwise method. In the backward selection method, initially, all predictors are included in the model. Then, it is tested if any of the predictor variables can be removed from the model without increasing the Akaike information criterion (AIC) [72]. If a variable can be removed, then after taking it away, the model is tested again on the remaining variables. This step is repeated until the removal of remaining variables results in no further reduction of the AIC value (for further detail see [37]).

For model evaluation, we split the data into 70% for training and 30% for testing. The performance measure or model goodness is reported for test data in terms of the Area Under the Receiver Operating Characteristic (ROC) Curve – also known as the AUC measure [11]. ROC is a probability curve, and the AUC value shows the extent to which the model can distinguish between classes. The higher the value of AUC, the better is the model at distinguishing between different classes, e.g., in the context of the current study, a higher AUC means that the model can distinguish well between the students who report a given TD versus those who do not report that TD. The value of AUC can range between 0 and 1. A model with an AUC of 0.5 works at the chance level and a model with an AUC value closer to 1 means that it has a good measure of separability between different classes.

5 Results

5.1 Relationship between Task Difficulties and Student Demographics

In terms of TDs across gender, males are more likely than females to perceive that the *Prediction* task is *easy*, $\chi^2(1, N = 180) = 3.94, p < .05$. Females, on the other hand, are more likely than males to perceive that the *Prediction* task is *hard*, $\chi^2(1, N = 180) = 9.35, p < .00$. The proportion of students who report that the *Prediction* task is *medium*, does not differ by gender $\chi^2(1, N = 180) = 0.72, p = .40$. Again, during the *Observation 1* task, males are marginally more likely than females $\chi^2(1, N = 146) = 3.14, p = .06$ to perceive that the task is *easy*; females are marginally more likely than males to perceive that the task is *hard*, $\chi^2(1, N = 146) = 3.38, p = .06$. Again, gender does not seem to influence students' perceptions when it comes to reporting that the TDs are *medium*, $\chi^2(1, N = 146) = 0.11, p = .74$.

In terms of ethnicity, during the *Observation 1* task, Hispanic/Latino are more likely ($p < .05$) than other students to report that the TD is *hard*. During the *Observation 2* task, students from the category 'others' are less likely than their peers to report that the

TD is *easy*, $\chi^2(2, N = 141) = 9.68, p = .01$. During *Observation 2* and *Explanation 1* task, Hispanic/Latino students are again more likely than their peers to report that the TDs are *hard* ($p < .05$ and $p < .10$ respectively). Finally, for the perceived difficulty *after-task*, the students who declare themselves as 'White' are more likely than other students to report that the overall difficulty is *easy*, $\chi^2(2, N = 179) = 5.86, p = .05$.

Lastly, TDs are compared for students across different age groups. During the *Prediction* task, the students between the age group 17-20 are less likely than other students to report that the TD is *medium*, $\chi^2(2, N = 163) = 7.37, p = .03$. While students aged 21-30 are likely to report that this task is *easy*, $\chi^2(2, N = 163) = 11.83, p = .00$, the students aged above 30 are likely ($p = .02$) to report that this task is *hard*. During, *Observation 1* task, students aged 17-20 are less likely than the other students to report that the TD is *easy*, $\chi^2(2, N = 163) = 7.37, p = .03$. During *Observation 2* task, students above the age of 30 are less likely $\chi^2(2, N = 130) = 6.99, p = .03$, to perceive that the task is *easy* and more likely ($p < .05$) to perceive that the task is *medium* than the other students. During *Explanation 1* task, students aged above 30 are again less likely $\chi^2(2, N = 72) = 5.12, p = 0.06$, to perceive that this task is *easy* than the other students; and students aged 17-20 are more likely ($p < .05$) to perceive that this task is *hard*. During the *Explanation 2* task, students aged above 30 are more likely ($p = .03$) to perceive that the task is *medium* than the other students. For the perceived difficulty *after-task*, age does not seem to influence students' perceptions.

Later, we investigated if there is an association between students' learning outcomes and their demographics. We find no significant differences e.g., when learning outcomes are compared between males and females $\chi^2(1, N = 161) = 0.08, p = 0.78$, when learning outcomes are compared for ethnicity $\chi^2(2, N = 161) = 3.37, p = 0.19$, or when learning outcomes are compared between students of different age groups $\chi^2(2, N = 161) = 1.16, p = 0.56$. It is interesting to note that the lack of difference in learning outcomes is despite the differential perception of difficulty.

5.2 Relationship between Task Difficulties and Learning Outcomes

A comparison of perceived difficulties, between the *high* achieving students and the *low* achieving students, is presented in **Figure-3**. The figure shows that most of the students in both groups perceived the tasks to be *easy*. However, when comparisons are made between the groups, it is found that the *high* achievers are more likely to perceive that the tasks are *medium* or moderately difficult than the *low* achievers – who seem to perceive that the tasks are either *hard* or *easy*. Overall, the proportion of students who respond during the *Explain-1* is the lowest, as this task is only available to the incorrect predicting students. Further, during the *Post POE* phase, many *high* achievers did not respond to the surveys. Therefore, the patterns during this task (where each TD category is more likely to be reported by the *low* achievers) differ from the overall trend.

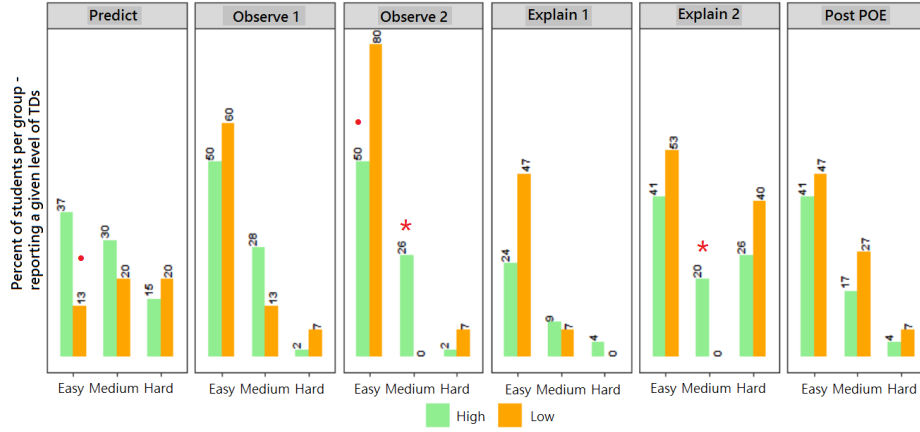


Figure-3: Comparison of TDs between the *high* and *low* achievers using Pearson's Chi-square test (or Fisher's exact test when the counts in the contingency table are less than 5). *High*-achievers tend to report *medium* TDs; in contrast, *low*-achievers tend to report the TDs as either *easy* or *hard*. Results are significant if $p\text{-value} < 0.05$ (*) and marginally significant if $p\text{-value} < 0.10$ (•).

5.3 Prior Knowledge and Learning Outcomes

To analyse if the *high* and the *low* achieving students differ in their prior knowledge, we compare their selected hypotheses from the *Prediction* task (see **Table-1**). Regardless of students' achievement, approximately 63% of students from each group choose an incorrect hypothesis. We further expect that many students may hold a common misconception that the bigger stars live longer. The results from **Table-1** indicate this to be the case with many students (~ 69%) in both groups endorsing this claim. This, seems to suggest that students in both groups had similar levels of prior knowledge before beginning the *POE* based tasks, and a majority of students in both groups held a common misconception.

Table-1: Comparison of the selected hypotheses for the *high* and the *low* achieving students during the *Prediction* task.

<i>Hypothesis</i>	<i>High achievers</i> ($n = 102$)	<i>Low achievers</i> ($n = 66$)
Correct	38 (37.6%)	24 (36.4%)
Incorrect	64 (62.4%)	42 (63.6%)
Common misconception	44 (68.8%)	29 (69.0%)
Other misconception	20 (31.2%)	13 (31.0%)

5.4 Analysis of Task Difficulty Transitions using *L*-Stat

The probability of TD transitions between the consecutive *POE* tasks reflects that students tend to report the *easy-easy* transition most frequently (see appendix **Table A2**). However, these probabilities are prone to error as they do not account for the base rate

of a given TD. To calculate the probability of TD transitions corrected for the base rate of a given TD, we use the likelihood statistic L -stat. **Table-2** presents an analysis of TD transitions or sequences in terms of D'Mello's L -statistic.

From this table, when self-transitions are analysed, the shift from *easy* → *easy* is not significantly more or less likely than chance. In contrast, the shift from *hard* → *hard* and from *medium* → *medium* are significantly less likely than chance. In terms of increasing TDs, a transition from the *easy* → *medium* is less likely than chance, from *easy* → *hard* is more likely than chance and from *medium* → *hard* is not different from chance level. Finally, in terms of decreasing TDs, the transitions from *hard* → *easy* and *medium* → *easy* are not different from the chance level; however, the *hard* → *medium* is more likely than chance.

Table-2. Sequences of TDs, using D'Mello's L -Statistic. L_{MEAN} in **bold** indicates the transition is more likely than chance and L_{MEAN} in *italics* indicates that the transition is less likely than chance.

Transitions		Descriptives			One-sample t -test		
from	to	N	L_{MEAN}	L_{SD}	T (df)	p -value	sig after BH correction
<i>easy</i>	<i>easy</i>	101	-0.01	0.63	-0.15 (100)	0.88	
	<i>medium</i>	121	<i>-0.44</i>	1.00	-4.85 (120)	<0.01	*
	<i>hard</i>	133	0.25	0.74	3.85 (132)	<0.01	*
<i>medium</i>	<i>easy</i>	130	-0.11	1.01	-1.24 (129)	0.22	
	<i>medium</i>	110	<i>-0.65</i>	1.27	-5.43 (109)	<0.01	*
	<i>hard</i>	138	-0.05	0.43	-1.48 (137)	0.14	
<i>hard</i>	<i>easy</i>	135	-0.08	0.70	-1.33 (134)	0.19	
	<i>medium</i>	139	0.14	0.47	3.36 (138)	<0.01	*
	<i>hard</i>	107	<i>-0.77</i>	1.28	-6.20 (106)	<0.01	*

5.5 Analysis of TD Transitions using Bi-grams

Next, we analyse students' perceived difficulty or TDs over consecutive tasks. We compare the students who report a given TD bigram sequence versus those who do NOT report that sequence. This analysis can assist in analysing how a sequence of TDs may impact students' learning outcomes (see **Table-3**). From this table, the performance is significantly low for the students who report the TD sequence *hard-hard* than those who do not report it. In contrast, the students who report the TD sequence *medium-medium* have significantly high scores than those who do not report it.

5.6 Task Difficulty Modelling

We develop separate models for each level of TD during each phase of the POE tasks. Detectors for TD easy are presented in

Table-4, for TD medium are presented in **Table-5**, and lastly, the detectors for TD hard are presented in **Error! Reference source not found.**. A definition of all the features considered for inclusion in these detectors is provided in appendix **Table A1**.

Table-3. TD sequences and their likely association with students' learning outcomes or their performance. Performance seems to be lower for the bigram sequence *hard-hard*, and it appears to be higher for the sequence *medium-medium*.

TD Bigram sequence	Bigram reporting students		T (59)	p-value	sig after BH correction
	Yes	No			
	Post-test (Mean \pm SD)	Post-test (Mean \pm SD)			
<i>easy-easy</i>	7.81 \pm 3.08	8.34 \pm 3.01	-1.12	0.26	
<i>easy-medium</i>	6.96 \pm 4.48	8.01 \pm 2.86	-1.34	0.18	
<i>easy-hard</i>	6.35 \pm 5.04	8.08 \pm 2.86	-1.86	0.06	
<i>medium-easy</i>	7.68 \pm 3.63	7.79 \pm 3.18	-0.15	0.88	
<i>medium-medium</i>	9.81 \pm 0.57	7.19 \pm 3.70	3.44	<0.01	*
<i>medium-hard</i>	8.67 \pm 1.70	7.66 \pm 3.60	0.62	0.54	
<i>hard-easy</i>	7.03 \pm 3.53	8.04 \pm 3.48	-1.22	0.22	
<i>hard-medium</i>	8.33 \pm 1.81	7.66 \pm 3.71	0.57	0.57	
<i>hard-hard</i>	6.35 \pm 5.58	8.18 \pm 2.49	-2.61	0.01	*

Regarding the performance, it appears that the detectors for TD *easy* perform better (mean AUC = 0.76) than the detectors for TD *hard* (mean AUC = 0.63), followed by the detectors for TD *medium* (mean AUC = 0.61). A reason for this is perhaps the varying sample size for each TD. We analysed the number of times each TD is reported and find that TD *easy* is reported (n = 311) more than TD *hard* (n = 122), followed by the TD *medium* (n = 114). Most of the task difficulty detectors are better than chance, but there is some room for improvement.

Overall, these detectors are developed using students' task-based interaction patterns as well as their perceived difficulties at a current task to predict their perceived difficulties at the following tasks. In addition to reporting the model performance in predicting the TDs at subsequent tasks, we also report the significant features that contribute the most in a model decision at each stage of the POE tasks.

Table-4. Detectors for TD *easy* are developed separately for each task of the POE cycle.

Detector performance is reported in terms of the AUC – Area Under the ROC Curve. Significant predictors for each detector are reported where (+) indicates positive predictor, and (-) indicates negative predictor.

Task	AUC	Sig. Features
<i>Prediction</i>	0.64	(-) Prior task time
<i>Observation1</i>	0.84	(-) <i>Prediction</i> incorrect hypothesis, (-) <i>Prediction</i> response length, (+) <i>Prediction Easy</i>
<i>Observation2</i>	0.71	(+) <i>Observation 1</i> attempts, (+) <i>Prediction Easy</i> , (+) <i>Observation 1 Easy</i>
<i>Explanation1</i>	0.96	(+) <i>Observation 2 Easy</i> , (+) <i>Observation 1 Easy</i> , (+) <i>Prediction Easy</i>
<i>Explanation2</i>	0.51	(-) <i>Observation 2</i> time, (+) <i>Explanation 1 Easy</i> , (+) <i>Prediction Easy</i>
<i>Post POE</i>	0.87	(+) <i>Explanation 2 Easy</i>

Table-5. Detectors for TD *medium* are developed separately for each task of the *POE* cycle. Model performance is reported in terms of the AUC – Area Under the ROC Curve. Significant predictors for each detector are reported where (+) indicates positive predictor, and (-) indicates negative predictor.

Task	AUC	Sig. Features
<i>Prediction</i>	0.63	(-) Prior task attempts, (+) Prior task time
<i>Observation1</i>	0.58	(+) <i>Prediction Medium</i>
<i>Observation2</i>	0.55	(+) <i>Prediction</i> response length, (+) <i>Observation 1 Medium</i>
<i>Explanation1</i>	0.50	(+) <i>Prediction</i> response length, (+) <i>Observation 2 Medium</i>
<i>Explanation2</i>	0.91	(+) <i>Explanation 1 Medium</i>
<i>Post POE</i>	0.51	(+) <i>Explanation 1 Medium</i>

Table-6. Detectors for TD *hard* developed separately for each stage of the *POE* cycle. Model performance is reported in terms of the AUC – Area Under the ROC Curve. Significant predictors for each detector are also reported where (+) indicates positive predictor, and (-) indicates negative predictor.

Task	AUC	Sig. Features
<i>Prediction</i>	0.48	(+) Prior task attempts
<i>Observation1</i>	0.77	(+) Prior task time, (+) <i>Prediction Attempts</i> , (+) <i>Prediction Hard</i>
<i>Observation2</i>	0.53	(+) Prior task time, (+) <i>Prediction Hard</i>
<i>Explanation1</i>	0.97	(+) <i>Observation 1 Hard</i>
<i>Explanation2</i>	0.50	(+) <i>Observation 1 Hard</i>
<i>Post POE</i>	0.51	(+) <i>Observation 1 time</i> , (+) <i>Explanation 1 Hard</i>

From these tables, it can be suggested that knowledge of prior tasks can guide the detection of perceived difficulties or TDs at the subsequent tasks, i.e., our calculations of TDs at later phases in the *POE* process can be improved by using the TDs from earlier phases. Moreover, there appears to be a proximity effect for each TD detector, where the perceived difficulties on a current task can be strong positive predictors of perceived difficulties at subsequent tasks.

6 Discussion

The goal of this study is to analyse students' perceptions of difficulties or TDs. For TD analysis, we use three labels, namely: *easy*, *medium* and *hard*. In terms of students' demographics, we find that females are more likely to report that the TDs are *hard* than males who are more likely to report that the TDs are *easy*. Similarly, Hispanic/Latino students are likely to say that the TDs are either *hard* or *medium*; by contrast, 'White' students are likely to report that the TDs are *easy*. In terms of age, students aged 21-30 are more likely to say that the TDs are *easy* than the students aged above 30 who are more likely to report that the TDs are either *hard* or *medium*. Despite the differences in perceived difficulties, when demographics are examined for learning outcomes, we find

no significant differences. Next, we discuss the research questions that are the focus of this study.

6.1 RQ1: Relation of task difficulties and students' learning outcomes

The first research question examines the relationship between students' TDs and their learning outcomes. From **Figure-3**, it is observed that during the *POE* sequence of tasks, the *low* achieving students mostly report the tasks as either *easy* or *hard*. For the *low* achievers who report that the tasks are *hard*, it could be that they struggled with the learning content, the environment or both. However, for the students who perceive that the tasks are *easy* and yet achieve poorer learning outcomes, a possible explanation for this could be their self-efficacy beliefs. Self-beliefs may influence students' performance [5, 6]. The students with unrealistic and overly optimistic opinions may have difficulty aligning their efforts with the desired performance levels, and that can subsequently deteriorate their performance [15, 17, 63].

Figure-3 further suggests that the *high* achieving students mostly report that the TDs are *medium*. A plausible explanation for this outcome is that students tend to engage more in the tasks that are perceived as moderately difficult than the tasks that are perceived too *easy* or too *hard* [8]. Therefore, for curricula design, the instructors should plan the tasks that are within the learners' zone of proximal development (ZPD) [85]. If learners are taught a skill that is within their ZPD, it can lead to better performance than when the skill is not [88]. In this regard, [22] suggests that subjects can perform at their optimal capabilities when they experience 'flow', which is likely to happen when their challenge regarding the tasks matches with their skills (confidence in this case).

It is important to mention that students' TDs from **Figure-3** seem to differ at the start of the *POE* tasks – the *Prediction* phase, where the *high* achieving students are more likely than the *low* achievers ($p\text{-value} < 0.10$), to indicate that the TDs are *easy*. This difference during the *Prediction* task is important as this task probes students' prior knowledge. Reporting this task *easy* can mean that these students have higher prior knowledge or higher confidence in prior knowledge which contributed to their performance [55, 56]. However, when the *high* and the *low* achieving students are assessed on the correctness of their hypotheses (see **Table-I**), we find that a comparable proportion of students across the two groups proposed an incorrect hypothesis. Further, when the nature of students' incorrect hypotheses across the two groups is compared, it seems that the *high* and the *low* achieving students hold misconceptions of a similar nature where most students seem to believe that the high mass stars live longer than the low mass stars. This may suggest that although the *high* achievers have similar levels of prior knowledge as the *low* achievers; they are more confident of their knowledge. In this regard, prior research suggests that students' confidence in the knowledge they hold is an important factor that can influence their learning and learning behaviours [55, 56].

Further, in a *POE* context, the *Observe* phase is crucial; it may provide valuable insights into students' prior held beliefs [33]. Confusion may be triggered for students who make incorrect *Predictions* [64]. Interestingly, there are more *low* achievers who make incorrect *Predictions*; yet the *low* achieving students are more likely to report that

this task is *easy* (p -value = 0.08). Thus, a knowledge of students' TDs at specific moments can help identify the students who may require interventions.

6.2 RQ2: Task difficulty transitions

The second research question analyses the transitions or sequences of TDs to assess how students' perceptions of difficulties or TDs change within the learning environment. Prior research on task-based instruction suggests that pedagogic tasks should be sequenced in increasing order of their demands or complexity [59, 75, 80]. For example, the cognition hypothesis suggests that a gradual increase in task complexity can prepare students for more advanced problems and can lead them to achieve better performance and development [73, 74, 75]. Within the current simulation environment, as the students progressed, the tasks became more complex (in terms of the required actions and activities). The impact of task complexity on TDs is presented in **Table-2**. From this table, the transition from *hard* \rightarrow *medium* is more likely than chance, while from *easy* \rightarrow *medium* is less likely than chance.

When the findings from RQ1 suggest that *medium* or moderate difficulties may lead to better learning outcomes, the results from RQ2 suggest that *harder* tasks are likely to be followed by moderate difficulties. This, then raises the question of how we can make all students experience difficulties of moderate level – should we intentionally make *harder* or complex tasks as they seem to precede TDs of *medium* level? Or should we make the follow-up tasks feel easier by comparison? This question may benefit from further studies where, e.g., we compare two groups, a treatment group may be offered less guidance from the system so that the tasks become more complex.

6.3 RQ3: Which sequence of task difficulties is better?

The third research question analyses the association between sequences of TDs and students' learning outcomes. Research on the sequential effects of TDs suggests that a learner's performance on a given task (regardless of whether the task is *easy* or *hard*) may be affected by the TDs on the preceding task [13, 78]. In their work, Schneider and Anderson [78] report that when an individual faces a *hard* task, a greater amount of cognitive resources may be allocated to it, and as they proceed to the next task, there may be a depletion in the available resources. Hence, the performance in the next task may be affected. To inspect this in more detail, we analyse the impact of TD sequences (over consecutive tasks) on students' learning outcomes. From **Table-3**, the students with perceived difficulty *hard* on two or more consecutive tasks are significantly more likely to have poorer learning outcomes than those who do not report such a transition. On the one hand, it could mean that these students are weak and therefore, perceive the tasks to be *hard*. On the other hand, it could also mean that perhaps there was a depletion of resources as students progressed from a *hard* task – which is in agreement with [78].

The next significant finding from **Table-3** is that the students who report *medium* difficulty on two or more consecutive tasks are likely to have better learning outcomes than the other students who do not report such a transition. What implications do these findings have for learning design? We find that *medium* TDs may lead to better learning

outcomes, and they often follow *hard* TDs. However, if tasks get too difficult for students, e.g., reporting *hard* on two or more consecutive tasks, then it can adversely affect students' performance. A knowledge of such perceptions of TDs, early on, may enable us to provide timely interventions to students.

6.4 RQ4: Modelling for task difficulties

The last research question aims to identify students' task-based interaction patterns that could be indicative of their perceived difficulties.

Detectors for TD Easy. In *Table-4*, detectors for TD *easy* are presented. Firstly, it appears that the students who find the current task to be *easy* are likely to report that the TD is *easy* on subsequent tasks, e.g., reporting *easy* during the *Prediction* task is a positive predictor for reporting *easy* during the *Observation 1* task. Similarly, reporting *easy* during the *Prediction* and *Observation 1* tasks can be predictive of reporting *easy* during the *Observation 2* task.

Interestingly, time on a current task seems to be a negative predictor for TD *easy* during a subsequent task, e.g., 'time on prior task' is found to be a negative predictor for TD *easy* during the *Prediction* task. Similarly, 'time on *Observation 2*' task is found to be a negative predictor for TD *easy* during the *Explanation 2* task. This can mean that the more time the students spend on a given task, the less likely they are to report that the subsequent task is *easy*, conversely speaking, the lesser the time the students spend on a given task, the more likely they are to report that the following task is *easy*.

We believe that the students who complete the tasks in a shorter time are less likely to reflect. As suggested by [87], the act of reflection can behaviourally be manifested in the form of pauses and rescanning which requires students to stop, think, ponder and possibly update the inconsistencies in their existing knowledge and the new information that they are exposed to. The act of reflection can thus be manifested in students' spending a longer time on tasks.

We also note that task attempts are a positive predictor of TD *easy* during subsequent tasks, e.g., the students who make more attempts during the *Observation 1* task are likely to perceive that the *Observation 2* task is *easy*. While from RQ1, we see that the *low* achieving students are likely to perceive the tasks as *easy*, what the behaviours from these detectors bring out is that perhaps these students are not reflecting on the tasks and are likely to complete the tasks in a shorter time, by making more attempts. Seemingly, these behaviours could be related to students' trial-and-error attempts where they are likely to game the system [2], which can ultimately result in these students' poorer learning outcomes.

Furthermore, from *Table-4*, we find that the length of students' textual reasoning during the *Prediction* task is a negative predictor for TD *easy* during the *Observation* task; suggesting that the students who write shorter texts are likely to perceive that the following task is *easy*. Students' behaviour of writing shorter texts for reasoning their choice of hypotheses, again seems to suggest that these students are reflecting less. For the texts written by students, [16] suggests that sentence length or word count is a

positive predictor of students' reflection behaviours and generally, the students who write and reflect more are likely to have better learning outcomes than their peers.

Overall, from these detectors, it appears that the students who perceive the tasks to be *easy* are less likely to engage with tasks probably because they feel that success can be obtained without deliberate effort. As a result, they reflect less and appear to be unable to develop a proper understanding of the concepts, as observed from their poorer learning outcomes.

From **Table-4**, it is also observed that the students who report that the *Prediction* task is *easy* are not only likely to report the immediately following task to be *easy* but also to report the other subsequent tasks as *easy*, e.g., reporting *easy* on *Prediction* task can be a positive predictor of reporting *easy* on *Observation 1*, *Observation 2*, *Explanation 1* and *Explanation 2* tasks. Additionally, the students who propose an incorrect hypothesis during the *Prediction* task, are less likely to report that the *Observation* task is *easy* than those who propose correct hypotheses. A reason for this behaviour can be that the incorrect predicting students during the *Observation* task may discover an inconsistency between their prior knowledge and the observations they are making. According to the cognitive disequilibrium theory [69, 70], when students are exposed to contradictions, anomalies, misconceptions or when they encounter a discrepant event where students' observations of a phenomenon are inconsistent with their expectations; cognitive disequilibrium is triggered. Cognitive disequilibrium is of critical importance in students' comprehension and learning processes [41]. Overall, these detectors highlight the importance of students' behaviours during the *Prediction* phase of a *POE* cycle.

Detectors for TD *Medium*. **Table-5** presents the detectors for TD *medium*. Firstly, for the *Prediction* task, those students who spent a long time on the prior task and who required fewer attempts to complete the prior task are likely to perceive that the *Prediction* task is *medium* or moderately difficult. What this behaviour could suggest is that these students are reflecting more on the prior task and thus, spent a longer time, which ultimately enabled them to complete the prior task in fewer attempts. This behaviour of reflection may then lead these students to report that the *Prediction* task is moderately difficult, where the challenge associated with the task may have matched students' skills. As a result, these students seemed to engage more during the *Prediction* task, and this is reflected in students' response length from the *Prediction* task. The students who appeared to be reflecting more, provided longer texts for hypotheses reasoning and then they were likely to perceive that the *Observation 2* and *Explanation 1* tasks are *medium*. Note that this relation was negative for perceived difficulty *easy*.

Seemingly, the students who reflect are likely to perceive the tasks as *medium* or moderately difficult or perhaps it is the *medium* difficulties that encourage students to engage and reflect more. This finding in terms of students' interaction patterns further explains the results from RQ1, where the *high* achieving students perceive the tasks to be *medium* or moderately difficult.

Lastly, like the detectors for TD *easy*, reporting *medium* on a current task is a positive predictor of reporting *medium* on the following task, e.g., those who report *medium* TD during the *Prediction* task are likely to report *medium* TD for *Observation 1* task. Those

who report *medium* TD on *Observation 1* task are likely to report *medium* TD on *Observation 2* task.

Overall, for TDs *easy* and *medium*, student behaviours and perceptions at the *Prediction* task seem to influence their behaviours at nearly all the *POE* based tasks. It suggests that the *Prediction* task (or in this case, the prior knowledge task) can strongly influence student perceptions at the following tasks.

Detectors for TD *Hard*. Lastly, we discuss the detectors for TD *hard*, as shown in **Table-6**. Generally, it appears that the students who spend a longer time on tasks and who require more attempts to complete the tasks are likely to perceive that the subsequent task is *hard*.

For example, students who make more attempts at the prior task are likely to report that the *Prediction* task is *hard*, and students who make more attempts during the *Prediction* task while spending a longer time at the prior task, are likely to report that the *Observation 1* task is *hard*. Similarly, spending a long time during the prior task and spending a long time during the *Observation 1* task are positive predictors for TD *hard* during *Observation 2* and *Post POE* tasks, respectively. Regarding the *Observation 1* task, we additionally see that the students who report TD to be *hard* during this task are likely to perceive that *Explain 1* and *Explain 2* tasks are *hard*.

These behaviours seem to suggest that for the *hard* TD detectors, students' interaction patterns during the *Observation 1* task are more important. *Observation 1* task is where students are first introduced to the stellar nursery simulator, and students are expected to learn to create and manipulate stars of varying mass. Finding this task *hard* could mean that these students struggled with the learning environment in general.

Lastly, like the previous detectors, the students who find the current task to be *hard* are likely to perceive that the subsequent tasks are *hard*. Overall, we believe that models or detectors of this nature can be used for student interventions or for discovery with model analyses.

7 Conclusion

In this study, we use task difficulties (TDs) as a factor of analysis. We find that when two groups of students start the *Predict-Observe-Explain (POE)* sequence of tasks with similar levels of prior knowledge and with misconceptions of a similar nature, one group can achieve better learning outcomes than the other group on the transfer task.

When the students in the two groups are inspected for perceived difficulties, we find some differences. Students who find the tasks to be *easy* or *hard* generally have poorer learning outcomes. However, if a task is perceived to be *easy*, and it is the prior knowledge task, it may lead to better learning outcomes; suggesting that these students may have higher confidence in prior knowledge.

Furthermore, in accordance with ZPD [85] and the flow theory [22], we find that TDs of *medium* level can lead to better performance. Researchers [34, 35] have acknowledged that only limited studies have investigated the role of students' TDs on their learning outcomes. We believe that this has an implication for AIED researchers

in that the TDs are based on students' subjective judgement of the tasks rather than the complexity of the tasks. This creates a possibility of individualised predictions of better paths to learning for each student.

In this study, we examine the effects of increasing as well as decreasing TDs on students' performance. An unexpected finding is that the students who find the current task *hard* are more likely to perceive that the following task is *medium* than the students who find the current task *easy*. This suggests that *hard* and challenging TDs have the potential to engage students and lead them to achieve better scores, and potentially influence their perceptions of the following tasks. However, when tasks become too *hard* (difficulty sustains over two or more tasks), then it can adversely affect students' performance. To control for the negative effects of TDs, one approach is to detect these difficulties early on so that personalised interventions are provided to enhance students' learning.

In this study, in addition to analysing students' TD sequences, we also analysed their learning processes or behaviours. We developed detectors for TDs to examine students' task-based interaction patterns that may lead to certain levels of TDs. When the detectors for TD *easy* are analysed, we find that TD *easy* is generally associated with students' spending lesser time and making more attempts on tasks, as well as providing shorter reasoning for justifying their choice of hypotheses. Generally, these students seem to game the system. For TDs *medium*, we find that usually the students who spend a longer time and make fewer attempts on tasks, as well as those who provide a longer text for reasoning their selected hypotheses than their peers, are likely to perceive that the tasks are moderately difficult. These behaviours are mostly indicative of a student who is reflecting more and engaging more with the tasks. For TD *hard*, it appears that the students who require more time and make more attempts to complete the tasks are likely to report that the TDs are *hard*. Overall, from these detectors, it was found that the TDs from the earlier phases of the *POE* process can guide our calculations of TDs at later phases. The reasoning for this could be that the earlier activities are a prerequisite of the later activities or that once students perceive a task to have a certain level of difficulty, students' feelings stick with them in the later stages of the *POE* process.

Overall, an understanding of how task difficulties manifest over time and how they impact students' learning outcomes is useful, especially when designing for real-time educational interventions, where the difficulty of the tasks can be optimised for the learners. It can also help in designing and sequencing the tasks, for the development of effective teaching strategies that can maximise student learning [58] and reduce undesirable behaviours such as gaming the system [2] and disengagement [39].

8 Future Work and Limitations

There are some limitations to the present analysis that need to be addressed in future. In this study, in addition to analysing students' behaviours (such as time on tasks, task attempts and textual responses) we have used self-reports as an external criterion, where students are asked to report their perceived difficulty regarding the tasks. Research suggests that self-reports can be influenced by social-desirability bias [54, 68], where

participants tend to report what seems favourable by society or by the researchers. Furthermore, self-reports can influence cognition in subtle ways [57]. Despite this, researchers in the AIED community have extensively used self-report instruments in their studies [18, 36, 38, 77].

Furthermore, in this work, TDs are analysed for a single *Predict-Observe-Explain* (POE) learning environment. In future, TDs can be analysed across different POE as well as non-POE learning environments to see if the findings of this study generalise across different contexts. This would allow a more in-depth understanding of student's TDs.

Lastly, to investigate how all students can be supported to experience *medium* or moderate levels of TDs, an experimental study can be designed. For example, we can compare two student groups; a treatment group may be offered lesser guidance from the system so that the tasks become more complex. This could provide further validity to the findings of this study where it was found that *harder* tasks are more likely to be followed by *medium* TDs, than the *easier* tasks.

Acknowledgements. We wish to thank Prof. Arial Anbar, Dr Lev Horodyskyj and Dr Chris Mead for providing us with the *Habitable Worlds* data for this research. We thank Dr Linda Corrin and Donia Malekian for the useful discussion on this work. We are also thankful to the anonymous reviewers for their time and valuable feedback on this paper. The quality of this manuscript has improved because of their insightful suggestions. This research is supported by the Research Training Program (RTP) Scholarship, Melbourne Research Scholarship and the Science of Learning Research Center (SLRC) top-up scholarship.

Note from authors. This paper is an expanded version of an earlier conference paper: Nawaz, S., Srivastava, N., Yu, J. H., Baker, R. S., Kennedy, G., & Bailey, J. (2020). Analysis of Task Difficulty Sequences in a Simulation-Based POE Environment. *Artificial Intelligence in Education: 21st International Conference, AIED 2020, Ifrane, Morocco, July 6–10, 2020, Proceedings, Part I*, 12163, 423–436. https://doi.org/10.1007/978-3-030-52237-7_34.

Table A1. Definition of various features for task difficulty detectors

Feature/Predictor	Definitions
Prior task time	Students' time spent on a prior task before the start of the POE sequence
Prior task attempts	Students' attempts on a prior task before the start the POE sequence of tasks
<i>Prediction</i> incorrect hypothesis	Binary variable suggesting the correctness of students' predictions where 1 = correct and 0 = incorrect
<i>Prediction</i> response length	Count of words that students write during the <i>Prediction</i> task for reasoning their choice of hypotheses.
POS	Count of positive terms from the texts that students write while reasoning their hypotheses.
NEG	Count of negative terms from the texts that students write while reasoning their hypotheses.
NEU	Count of neutral terms (including articles and pronouns) from the texts that students write while reasoning their hypotheses.
<i>Prediction</i> time	Time spent during the <i>Prediction</i> task
<i>Prediction</i> attempts	Attempts made during the <i>Prediction</i> task
<i>Prediction easy</i> , <i>Prediction medium</i> , <i>Prediction hard</i>	These features represent students' perceived difficulty during the <i>Prediction</i> task. For each detector, the perceived difficulty is binary, e.g., when detectors for TD <i>easy</i> are considered then the feature <i>Prediction easy</i> could either be one (1) which means that the TD is reported <i>easy</i> or zero (0) which implies that the reported TD is not <i>easy</i> .
<i>Observation 1</i> time	Time spent during the <i>Observation 1</i> task
<i>Observation 1</i> attempts	Attempts made during the <i>Observation 1</i> task
<i>Observation 1 easy</i> , <i>Observation 1 medium</i> , <i>Observation 1 hard</i>	These features represent students' perceived difficulties during the <i>Observation 1</i> task. Like before, for each detector, the perceived difficulty is binary, e.g., when detectors for TD <i>medium</i> are considered then the <i>Observation 1 medium</i> could either be one (1) that implies the reported TD is <i>medium</i> or zero (0) which means the perceived difficulty is not <i>medium</i> .
<i>Observation 2</i> time	Time spent during the <i>Observation 2</i> task
<i>Observation 2</i> attempts	Attempts made during the <i>Observation 2</i> task
<i>Observation 2 easy</i> , <i>Observation 2 medium</i> , <i>Observation 2 hard</i>	Like before, these features represent students' perceived difficulties during the <i>Observation 2</i> task. For each detector, the perceived difficulty is binary, e.g., when detectors for TD <i>hard</i> are considered then the <i>Observation 2 hard</i> could either be one (1) which implies the reported TD is <i>hard</i> or zero (0) which means the perceived difficulty is not <i>hard</i> .
<i>Explanation 1 easy</i> , <i>Explanation 1 medium</i> , <i>Explanation 1 hard</i>	Similar to the above
<i>Explanation 2 easy</i> , <i>Explanation 2 medium</i> , <i>Explanation 2 hard</i>	Similar to the above

Table A2. Probability of TD transitions between consecutive *POE* tasks.

<i>Transitions</i>		<i>POE sequence of tasks</i>				
<i>from</i>	<i>to</i>	<i>Pred-Obs1</i>	<i>Obs1-Obs2</i>	<i>Obs2-Exp1</i>	<i>Exp1-Exp2</i>	<i>Exp2-postPOE</i>
<i>easy</i>	<i>easy</i>	0.32	0.49	0.51	0.31	0.37
	<i>medium</i>	0.09	0.04	0.02	0.03	0.06
	<i>hard</i>	0.02	0.01	0.01	0.17	0.02
<i>medium</i>	<i>easy</i>	0.11	0.13	0.02	0.01	0.06
	<i>medium</i>	0.14	0.10	0.09	0.06	0.05
	<i>hard</i>	0.03	0.03	0.03	0.03	0.03
<i>hard</i>	<i>easy</i>	0.09	0.09	0.01	0.01	0.08
	<i>medium</i>	0.05	0.05	0.01	0.00	0.14
	<i>hard</i>	0.15	0.07	0.04	0.06	0.18

References

1. Arroyo I, Woolf BP, Cooper DG et al. (2011) The impact of animated pedagogical agents on girls' and boys' emotions, attitudes, behaviors and learning. In: International Conference on Advanced Learning Technologies (ICALT). IEEE, Athens, GA, p 506-510
2. Baker R, Corbett AT, Koedinger KR et al. (2004) Off-task behavior in the cognitive tutor classroom: when students "game The system". In: SIGCHI Conference on Human Factors in Computing Systems. ACM, p 383-390
3. Baker R, D'mello S, Rodrigo MMT et al. (2010) Better to be frustrated than bored: The incidence, persistence, and impact of learners' cognitive-affective states during interactions with three different computer-based learning environments. *International Journal of Human-Computer Studies* 68:223-241
4. Baker RS, Ogan AE, Madaio M et al. (2019) Culture in Computer-Based Learning Systems: Challenges and Opportunities. *Computer-Based Learning in Context* 1:1-13
5. Bandura A (1997) Self-efficacy: The exercise of control. In: Freeman, New York
6. Bandura A (1977) *Social Learning Theory*. Prentice Hall, Englewood Cliffs, NJ
7. Bell AW, Brekke G, Swan M (1987) Misconceptions, conflict and discussion in the teaching of graphical interpretation. In: Novak J (ed) *Proceedings of the Second International Seminar: Misconceptions and Educational Strategies in Science and Mathematics*. Cornell University, Ithaca, N.Y., p 46-48
8. Belmont JM, Mitchell DW (1987) The general strategy hypothesis as applied to cognitive theory in mental retardation. *Intelligence* 11:91- 105
9. Berkson J (1944) Application of the logistic function to Bio-Assay. *Journal of the American Statistical Association* 39:357-365
10. Bjork RA (2013) Desirable difficulties perspective on learning. *Encyclopedia of the Mind* 4:134-146
11. Bradley AP (1997) The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognition* 30:1145-1159

12. Bunderson VC, Inouye DK, Olsen JB (1989) The four generations of computerized educational measurement. In: Linn RL (ed) Educational measurement. Macmillan, New York, NY
13. Campbell DJ (2008) Subtraction by addition. *Memory & Cognition* 36:1094–1102
14. Campbell DJ (1988) Task complexity: A review and analysis. *Academy of Management Review* 13:40-52
15. Carpentar VL, Friar S, Lipe MG (1993) Evidence on the performance of accounting students: Race, gender and expectations. *Issues in Accounting Education* 8:1-17
16. Chen Y, Yu B, Zhang X et al. (2016) Topic modeling for evaluating students' reflective writing: a case study of pre-service teachers' journals. In: Gašević D, Lynch G, Dawson S, Drachsler H, Rosé CP (eds) *Learning Analytics and Knowledge (LAK)*. Association for Computing Machinery (ACM), Edinburgh, UK, p 1-5
17. Christensen TE, Fogarty TJ, Wallace WA (2002) The association between the directional accuracy of self-efficacy and accounting course performance. *Issues in Accounting Education* 17:1-26
18. Colthorpe K, Zimbardi K, Ainscough L et al. (2015) Know Thy Student! Combining Learning Analytics and Critical Reflections to Increase Understanding of Students' Self-Regulated Learning in an Authentic Setting. *Journal of Learning Analytics* 2:134-155
19. Coştu B, Ayas A, Niaz M (2012) Investigating the effectiveness of a POE-based teaching activity on students' understanding of condensation. *Instructional Science* 40:47-67
20. Craig S, Graesser A, Sullins J et al. (2004) Affect and learning: an exploratory look into the role of affect in learning with AutoTutor. *Journal of Educational Media* 29:241-250
21. Csikszentmihalyi M (2000) *Beyond boredom and anxiety*. Jossey-Bass
22. Csikszentmihalyi M (1997) *Finding flow: The psychology of engagement with everyday life*.
23. Csikszentmihalyi M (1979) The flow experience. *Consciousness: Brain and states of awareness and mysticism*:63-67
24. Csikszentmihalyi M (1990) *Flow: The psychology of optimal experience*. Harper Perennial, New York
25. D'mello S, Graesser A (2014) Confusion and its dynamics during device comprehension with breakdown scenarios. *Acta psychologica* 151:106-116
26. D'mello S, Graesser A (2010) Modeling cognitive-affective dynamics with Hidden Markov Models. In: Annual meeting of the Cognitive Science Society. p 2721-2726
27. D'mello S, Person N, Lehman B (2009) Antecedent-consequent relationships and cyclical patterns between affective states and problem solving outcomes. In: *Artificial Intelligence in Education (AIED)*. p 57-64
28. D'mello S, Taylor RS, Graesser A (2007) Monitoring affective trajectories during complex learning. In: Annual Meeting of the Cognitive Science Society. p 203-208
29. D'mello S, Graesser A (2012) Dynamics of affective states during complex learning. *Learning and Instruction* 22:145-157
30. Dalziel J (2010) *Practical eTeaching strategies for predict – observe – explain, problem-based learning and role plays*. LAMS International, Sydney, Australia

31. Deloache JS, Cassidy DJ, Brown AL (1985) Precursors of mnemonic strategies in very young children's memory. *Child Development* 56:125-137
32. Dowell NMM, Graesser A (2014) Modeling learners' cognitive, affective, and social processes through language and discourse. *Journal of Learning Analytics* 1:183-186
33. Driver R (1983) *The pupil as scientist?* Open University Press, UK
34. Eccles JS, Adler TF, Futterman R et al. (1983) Expectancies, values and academic behaviors. In: Spence JT (ed) *Achievement and achievement motives*. W.H. Freeman, San Francisco, p 75-146
35. Eccles JS, Wigfield A (2002) Motivational beliefs, values, and goals. *Annual Review of Psychology* 53:109-132
36. Ellis RA, Han F, A. P (2017) Improving Learning Analytics – Combining Observational and Self-Report Data on Student Learning. *Educational Technology & Society* 20:158-169
37. Field A, Miles J, Field Z (2012) *Discovering statistics using R*. Sage publications
38. Gašević D, Jovanovic J, Pardo A et al. (2017) Detecting learning strategies with analytics: Links with self-reported measures and academic performance. *Journal of Learning Analytics* 4:113-128
39. Gobert JD, Baker R, Wixon MB (2015) Operationalizing and detecting disengagement within online science microworlds. *Educational Psychologist* 50:43-57
40. Gorsky P, Finegold M (1994) The role of anomaly and of cognitive dissonance in restructuring students' concepts of force. *Instructional Science* 22:75-90
41. Graesser A, Lu S, Olde BA et al. (2005) Question asking and eye tracking during cognitive disequilibrium: Comprehending illustrated texts on devices when the devices break down. *Memory & Cognition* 33:1235-1247
42. Greiff S, Wüstenberg S, Csapó B et al. (2014) Domain-general problem solving skills and education in the 21st century. *Educational Research Review* 13:74-83
43. Guadagnoli MA, Lee TD (2004) Challenge point: A framework for conceptualizing the effects of various practice conditions in motor learning. *Journal of Motor Behavior* 36:212-224
44. Gunstone R, White R (1980) A matter of gravity. *Research in Science Education* 10:35-44
45. Hom HL, Maxwell FR (1983) The impact of task difficulty expectations on intrinsic motivation. *Motivation and Emotion* 7:19-24
46. Horodyskyj LB, Mead C, Belinson Z et al. (2018) Habitable Worlds: Delivering on the Promises of Online Education. *Astrobiology* 18:86-99
47. Kapur M, Bielaczyc K (2012) Designing for productive failure. *The Journal of the Learning Sciences* 21:45-83
48. Kapur M, Rummel N (2012) Productive failure in learning and problem solving. *Instructional Science* 40:645-650
49. Karumbaiah S, Andres JMaL, Botelho AF et al. (2018) The implications of a subtle difference in the calculation of affect dynamics. In: *International Conference for Computers in Education*.
50. Karumbaiah S, Baker R, Ocumpaugh J (2019) The case of self-transitions in affective dynamics. In: *Artificial Intelligence in Education (AIED)*. p 172-181

51. Karumbaiah S, Ocumpaugh J, Baker R (2019) The Influence of School Demographics on the Relationship between Students' Help-Seeking Behavior and Performance and Motivational Measures. In: Lynch CF, Merceron A, Desmarais M, Nkambou R (eds) International Conference on Educational Data Mining (EDM). International Educational Data Mining Society, p 99-108
52. Kennedy G, Lodge J (2016) All roads lead to Rome: Tracking students' affect as they overcome misconceptions. In: 33rd International Conference of Innovation, Practice and Research in the Use of Educational Technologies in Tertiary Education (ASCILITE). Adelaide, AU, p 317
53. Kibirige I, Osodo J, Tlala KM (2014) The effect of Predict-Observe-Explain strategy on learners' misconceptions about dissolved salts. *Mediterranean Journal of Social Sciences* 5
54. Krosnick JA (1999) Survey research. *Annual Review of Psychology* 50:537-567
55. Kulhavy RW (1977) Feedback in written instruction. *Review of educational research* 47:211-232
56. Kulhavy RW, Yekovich FR, Dyer JW (1976) Feedback and response confidence. *Journal of Educational Psychology* 68:522-528
57. Lazarus RS (1991) Cognition and motivation in emotion. *American Psychologist* 46:352-367
58. Li W, Lee A, Solmon M (2007) The role of perceptions of task difficulty in relation to self-perceptions of ability, intrinsic value, attainment value, and performance. *European Physical Education Review* 13:301-318
59. Long MH, Crookes G (1992) Three approaches to task-based syllabus design. *TESOL quarterly* 26:27-56
60. Mangos PM, Steele-Johnson D (2001) The role of subjective task complexity in goal orientation, self-efficacy, and performance relations. *Human Performance* 14:169-185
61. Mayfield E, Madaio M, Prabhume S et al. (2019) Equity beyond bias in language technologies for education. In: In Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications. Association for Computational Linguistics, p 444-460
62. Maynard DC, Hakel MD (1997) Effects of Objective and Subjective Task Complexity on Performance. *Human Performance* 10:303-330
63. Mooi TL (2006) Self-efficacy and student performance in an accounting course. *Journal of Financial Reporting and Accounting* 4:129-146
64. Nawaz S, Kennedy G, Bailey J et al. (2020) Moments of confusion in simulation-based learning environments. *Journal of Learning Analytics* 7:118-137
65. Nawaz S, Kennedy G, Bailey J et al. (2018) Struggle town? Developing profiles of student confusion in simulation-based learning environments. In: M. Campbell, J. Willems, C. Adachi, D. Blake, I. Doherty, S. Krishnan, S. Macfarlane, L. Ngo, M. O'Donnell, S. Palmer, L. Riddell, I. Story, Suri H, Tai J (eds) 35th International Conference on Innovation, Practice and Research in the Use of Educational Technologies in Tertiary Education, ASCILITE 2018. Deakin University, Geelong, Australia, p 224-233

66. Ocumpaugh J, Baker R, Gowda S et al. (2014) Population validity for Educational Data Mining models: A case study in affect detection. *British Journal of Educational Technology* 45:487-501
67. Paquette L, Ocumpaugh J, Li Z et al. (2020) Who's learning? Using demographics in EDM research. *Journal of Educational Data Mining* 12:1-30
68. Paulhus DL (1991) Measurement and control of response bias. In: J.P. Robinson, P.R. Shaver, Wrightsman LS (eds) *Measures of personality and social psychological attitudes*. Academic Press, San Diego, CA, p 17-59
69. Piaget J (1985) *The equilibration of cognitive structures: The central problem of intellectual development*. University of Chicago Press
70. Piaget J (1952) *The origins of intelligence in children*. International University Press, New York
71. Pintrich PR, Schunk DH (2002) *Motivation in education: Theory, research, and applications*. Prentice Hall
72. Posada D, Buckley TR (2004) Model selection and model averaging in Phylogenetics: Advantages of Akaike Information Criterion and Bayesian approaches over likelihood ratio tests. *Systematic Biology* 53:793-808
73. Robinson P (2005) Cognitive complexity and task sequencing: A review of studies in a Componential Framework for second language task design. *International Review of Applied Linguistics in Language Teaching* 43:1-33
74. Robinson P (2001) Task complexity, cognitive resources and syllabus design: A triadic framework for examining task influences on SLA. In: Robinson P (ed) *Cognition and second language instruction*. Cambridge University Press, New York, p 185-316
75. Robinson P (2001) Task complexity, task difficulty, and task production: Exploring interactions in a componential framework. *Applied Linguistics* 22:27-57
76. Rodrigo MMT, Baker R, Agapito J et al. (2012) The effects of an interactive software agent on student affective dynamics while using an intelligent tutoring system. In: *IEEE Transactions on Affective Computing*. p 224–236
77. Rodriguez F, Yu R, Park J et al. (2019) Utilizing learning analytics to map students' self-reported study strategies to click behaviors in STEM courses. In: *International conference on learning analytics & knowledge (LAK)*. Tempe, AZ, USA, p 456-460
78. Schneider DW, Anderson JR (2010) Asymmetric switch costs as sequential difficulty effects. *The Quarterly Journal of Experimental Psychology* 63:1873-1894
79. Shute VJ, D'mello S, Baker R et al. (2015) Modeling how incoming knowledge, persistence, affective states, and in-game progress influence student learning from an educational game. *Computers & Education* 86:224-235
80. Skehan P (1998) *A cognitive approach to language learning*. Oxford University Press, Oxford
81. Sreerekha S, Arun RR, Sankar S (2016) Effect of Predict-Observe-Explain strategy on achievement in chemistry of secondary school students. *International Journal of Education & Teaching Analytics* 1
82. Stephanou G, Kariotoglou P, Dinas KD (2011) University students' emotions in lectures: The effect of competence beliefs, value beliefs and perceived task-difficulty, and the impact on academic performance. *International Journal of Learning* 18:45-72

83. Tao PK, Gunstone RF (1999) The process of conceptual change in force and motion during computer-supported physics instruction. *Journal of Research in Science Teaching* 36:859-882
84. Vosniadou S (1994) Capturing and modeling the process of conceptual change. *Learning and Instruction* 4:45-69
85. Vygotsky LS (1978) *Mind and society: The development of higher mental processes*. Harvard University Press, Cambridge, MA
86. White R, Gunstone R (1992) *Probing understanding*. Routledge
87. Yancey KB (1998) *Reflection in the Writing Classroom*. Utah State University
88. Zou X, Ma W, Ma Z et al. (2019) Towards helping teachers select optimal content for students. In: Isotani S, Millán E, Ogan A, Hastings P, McLaren B, Luckin R (eds) *International Conference on Artificial Intelligence in Education (AIED)*. Springer, Cham, Chicago, IL, p 413-417