Analysis of Task Difficulty Sequences in a Simulation-based POE Environment

Sadia Nawaz[1(✉)], Namrata Srivastava[1], Ji Hyun Yu[2], Ryan S. Baker[3], Gregor Kennedy[1] and James Bailey[1]

[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
[2] University of Pennsylvania, Philadelphia, PA 19104, United States

Abstract. Task difficulty (TD) reflects students' subjective judgement on the complexity of a task. We examine the task difficulty sequence data of 236 undergraduate students in a simulation-based *Predict-Observe-Explain* environment. The findings suggest that if students perceive the TDs as *easy* or *hard*, it may lead to poorer learning outcomes, while the *medium* or moderate TDs may result in better learning outcomes. In terms of TD transitions, difficulty level *hard* followed by a *hard* may lead to poorer learning outcomes. By contrast, difficulty level *medium* followed by a *medium* may lead to better learning outcomes.

Understanding 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 could be optimised for students. It can also help in designing and sequencing the tasks for the development of effective teaching strategies that can maximize students' learning.

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

1 Introduction

Students' perceptions of tasks can influence their learning behaviours [4, 6]. 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 [28, 49]. This, then, engenders the question: how can instructors design optimal learning conditions where students get challenged but feel confident in accomplishing the task? To address this question, we analyse the relation of task difficulties (TDs) with students' learning outcomes. Further, we observe how TDs vary in a simulation-based learning environment (e.g., is it more probable for TDs to transition from *easy* to *hard* or vice-versa). Lastly, we assess whether students' sequences of TDs can be indicative of their learning outcomes.

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 [18, 19, 21, 22, 36, 37]. Compared to a traditional classroom environment, a benefit of

analyzing TDs in a digital setting is that students can receive just-in-time support. For instance, the level of TDs can be adjusted by the instructors to match student's level of understanding or individual students may also choose and change the level of TD in a self-controlled setting [3, 25, 30, 62]. We believe that a better understanding of students' TDs will enable interventions to improve students' learning [1, 53, 55] and reduce undesirable behaviours such as gaming the system [2] and disengagement [29].

2 Related Work

Task complexity and task difficulty (TD) are often used interchangeably. However, they are two different constructs [51, 52]. Task difficulty refers to a person's subjective judgment on the complexity of a task, whilst task complexity represents the characteristics or cognitive demands of a task [9].

Different learners can perceive the same tasks differently [9]. Researchers have shown that TDs can influence students' motivation [32] and self-regulation [4]. TDs can also affect problem-solving strategies and tactics. For example, DeLoache, Cassidy and Brown [24] 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 [6]. Vygotsky [60] 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 [14, 58] 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 [15, 16]. 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 [44, 45], yet [7] states that 'certain difficulties can enhance learning'. Several studies have indicated that students can learn from challenges that lead them to identify and articulate their current views, examine their ideas and clarify their misconceptions [34, 35]. To sum up, we investigate the following questions in this paper:

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?

3 Learning Environment

3.1 Predict Observe Explain (POE) Simulations

This study is built on an underlying educational framework known as the *Predict-Observe-Explain* (POE) paradigm [61]. POE is a three-phase, iterative design [23].

1. During *Prediction*, students formulate a hypothesis. They are often asked to provide the reasons as to why they committed to it.
2. During *Observation*, students test their hypothesis by changing parameters or variables in a simulation. They can then see the effects of their manipulations. This phase is especially crucial for those who make incorrect hypotheses, as they can see a mismatch between their predictions and observations [26].
3. During the *Explanation* phase, clarifications are provided to students detailing the relationship between variables or parameters that represent the conceptual phenomenon under investigation. This phase assists students to reconcile any discrepancies between what they predicted and what they observed in the simulation [31].

POEs can be applied in face-to-face, online and computer lab contexts [13]. They can promote student discussion [61], probe into their prior knowledge and help them update prior conceptions [12, 39, 59]. POE learning designs can make digital environments more engaging [39, 57]. Recently, POE environments have been analysed to examine students' affective experience [38] and their behaviours relating to struggle and confusion [47, 48].

To the best of our knowledge, TDs have not yet been investigated within POE based environments. Understanding how TDs manifest over time and how they impact students' learning outcomes is useful, especially when designing for real-time 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 [33]. 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 platform[1], which records moment by moment activity of students. This adaptive learning environment allows the provision of feedback based on students' responses or lack of responses. This course is offered to non-science major 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 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,

[1] https://www.smartsparrow.com/research/

students follow the prescribed sequence of tasks or activities. Occasionally, however, there is pathways adaptivity for the remediation of students who make errors. Further, the students cannot proceed onto the next tasks unless the current task is completed.

3.3 Tasks Description

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

- *Prediction*: Students need to select a hypothesis from five possible choices regarding the relationship between stellar mass and lifespan. Then, they need to report their reasons (through free text) for selecting that 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. 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 the 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 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) [48].
- *Post POE*: At the final stage, 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 in the nuclear fusion process which fuels the burning of stars.

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.

3.5 Measures

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 is ten; with each incorrect attempt, students are penalized by two 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 scale: from 1 (not at all) to 6 (extremely). 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 sequence, students can again report their confidence and challenge on a 6-point scale when 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 TD during the *Prediction* task, 151 and 146 during the *Observe-1* and *Observe-2* tasks respectively, 74 and 146 during the *Explain-1* and *Explain-2* tasks. Lastly, 185 students report their perceived TD *after-task*.

4 Data Pre-processing

4.1 Levels of Task Difficulty

For analyzing the TD dynamics, 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 following (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 the reported challenge

Note that our TD labels match with Csikszentmihalyi's flow theory [17]. While the flow theory reports on students' affects in terms of their challenge and skills; we use these measures (challenge and confidence) to infer students' perceptions of difficulties.

4.2 Task Difficulties and Learning Outcomes

Learning outcomes reflect students' scores at the transfer task. The maximum achievable score is 10, and for each repeated attempt at this task two points are deducted. *High* achieving students are those who score 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.

4.3 Task Difficulty Sequences

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

Calculating *L*-stat. After obtaining students' TD sequences, we compute the likelihoods of transitions between any two possible states using the transition metric *L* [21], 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* (*difficulty*$_t$ → *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:

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

The value of *L* may vary from $-\infty$ to 1. For a given transition, if $L \approx 0$, we say that the transition occurs at chance level, if $L > 0$, we say that the transition is more likely than chance. Finally, if $L < 0$ then the transition is less likely than chance [20].

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 [36].

Generating bi-gram sequences. We process students' TD sequences to generate TD bigrams. We only consider the students who respond to all task-based surveys and who also attempt the transfer task – there are 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 the analysis also involves multiple comparisons, BH post-hoc correction is applied.

5 Results

5.1 Task Difficulties across Different Achievement Levels

A comparison of perceived difficulties, between the *high* achieving students and the *low* achieving students, is presented in Fig. 1. The high achievers are more likely to perceive the tasks as *medium* or moderately difficult than the low achievers – who seem to perceive the tasks as 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 of the 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.

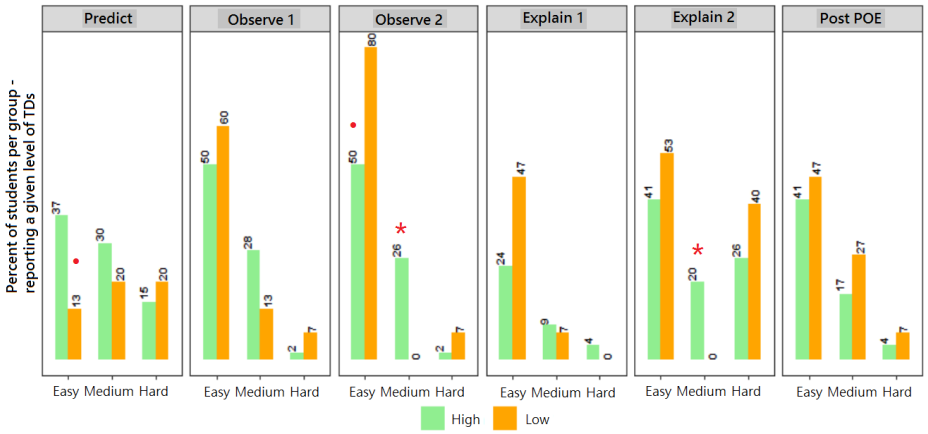

Fig. 1: Comparison of TDs between the high and low achievers using Pearson's Chi-square test (or the Fisher's exact test when the counts in the contingency table are less than 5). High-achievers tend to report *medium* TDs; in contrast, low-achievers tend to report the TDs as either *easy* or *hard*. Results are significant if p-value<0.05 (*) and marginally significant if p-value<0.10 (·).

5.2 Analysis of Task Difficulty Sequences

Table 1 presents the TD dynamics in terms of D'Mello's *L* statistic. For self-transitions, 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 *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 from *medium* → *easy* are not different from chance level, however, from *hard* → *medium* is more likely than chance.

Table 1. Dynamics of TDs, using D'Mello's *L*-Statistic. L_{MEAN} in **bold** indicates the transition is more likely 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*
easy	*easy*	101	-0.01	0.63	-0.15 (100)	0.88	
	medium	121	*-0.44*	1.00	-4.85 (120)	<0.01	*
	hard	133	**0.25**	0.74	3.85 (132)	<0.01	*
medium	*easy*	130	-0.11	1.01	-1.24 (129)	0.22	
	medium	110	*-0.65*	1.27	-5.43 (109)	<0.01	*
	hard	138	-0.05	0.43	-1.48 (137)	0.14	
hard	*easy*	135	-0.08	0.70	-1.33 (134)	0.19	
	medium	139	**0.14**	0.47	3.36 (138)	<0.01	*
	hard	107	*-0.77*	1.28	-6.20 (106)	<0.01	*

5.3 Analysis of Bi-gram Sequences

Next, we analyze students' perceived difficulty over consecutive tasks. We compare the students who report a given bigram sequence versus those who do NOT report it. This analysis can assist in analyzing how a sequence of TDs may impact students' post-test performance (see Table 2). 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.

Table 2. TD sequences and their likely association with students' 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 ± SD)	*Post-test (Mean ± SD)*			
easy-easy	7.81 ± 3.08	8.34 ± 3.01	-1.12	0.26	
easy-medium	6.96 ± 4.48	8.01 ± 2.86	-1.34	0.18	
easy-hard	6.35 ± 5.04	8.08 ± 2.86	-1.86	0.06	
medium-easy	7.68 ± 3.63	7.79 ± 3.18	-0.15	0.88	
medium-medium	9.81 ± 0.57	7.19 ± 3.70	3.44	<0.01	*
medium-hard	8.67 ± 1.70	7.66 ± 3.60	0.62	0.54	
hard-easy	7.03 ± 3.53	8.04 ± 3.48	-1.22	0.22	
hard-medium	8.33 ± 1.81	7.66 ± 3.71	0.57	0.57	
hard-hard	6.35 ± 5.58	8.18 ± 2.49	-2.61	0.01	*

6 Discussion

The goal of this study is to analyse the perceptions of difficulties or TDs. For analysis, we use three labels namely: *easy*, *medium* and *hard*.

RQ1. The first research question examines the relationship between students' TDs and their learning outcomes. From Fig. 1 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 the tasks as *hard*, it could be that they struggled with the learning content, the environment or both. However, for the students who perceive the tasks as *easy* and yet achieve poorer learning outcomes, a possible explanation for this could be their self-efficacy beliefs. Self-beliefs may influence students' performance [4, 5]. 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 [10, 11, 46].

Fig. 1 further suggests that the high achieving students mostly report the TDs as *medium*. A plausible explanation for this outcome is that students tend to engage more in the tasks that are perceived moderately difficult than the tasks that are perceived too *easy* or too *hard* [6]. Therefore, for curricula design, the instructors should plan the

tasks that are within the learners' zone of proximal development (ZPD) [60]. If learners are taught a skill that is within their ZPD, it can lead to better performance than when the skill is not [62]. In this regard, [15] 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 Fig. 1 seem to differ at the start of the POE tasks – the *Prediction* phase, where the high achieving students are more likely (*p-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* could mean that these students have higher prior knowledge or higher confidence in prior knowledge which contributed to their performance [40, 41].

Further, in a POE context, the *Observe* phase is crucial, it may provide valuable insights into students' prior held beliefs [26]. Confusion may be triggered for students who make incorrect *Predictions* [47]. Interestingly, there were more low achievers who made incorrect *Predictions*; yet the low achieving students were more likely to report this task as *easy* (*p-value* = 0.08). Thus, knowledge of students' TDs at specific moments can help identify the students who require interventions.

RQ2. The second research question analyses the dynamics of TDs – how students' perceptions of difficulties change within this environment. Prior research on task-based instruction suggests that pedagogic tasks should be sequenced in increasing order of their demands or complexity [43, 52, 56]. 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 [50, 51, 52]. 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 1. From this table, the transition from *hard* → *medium* is more likely than chance, while from *easy* → *medium* is less likely than chance.

When the findings from RQ1 suggest that *medium* or moderate difficulty may lead to better learning outcomes, the results from RQ2 suggest that *harder* tasks are likely to be followed by moderate difficulty. 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? We believe that 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.

RQ3. The last 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 [8, 54]. In their work, Schneider and Anderson [54] 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 2, the students

with perceived difficulty of *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 as *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 [54].

The next significant finding from Table 2 is that the students who report *medium* difficulty on two or more consecutive tasks are likely to have better learning outcomes than other students. 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.

7 Conclusion

In this study, we use task difficulties (TDs) as a factor of analysis. Researchers [27, 28] have acknowledged that only limited studies have investigated the role of students' TDs on their learning outcomes. We examine the effects of increasing as well as decreasing TDs on students' performance. Students who find the tasks *easy* or *hard* generally have poorer learning outcomes. However, if a task is perceived *easy* and it is the prior knowledge task, it may lead to better learning outcomes. Furthermore, in accordance with ZPD [60] and the flow theory [15], we find that TDs of *medium* level can lead to better performance. An implication for AIED researchers is that, TDs are based on students' subjective judgement of the task rather than task complexity. This creates a possibility of individualized predictions of better paths to learning for each student.

An unexpected finding was that the students who find the current task to be *hard* are more likely to perceive the following task as *medium* than the students who find the current task to be *easy*. This suggests that *hard* and challenging TDs have the potential to engage students and lead them to achieve better scores, as well as potentially influencing perception of 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.

A potential future direction for this work could be the analysis of students' learning behaviours to see how some students who find the current task to be *hard* can overcome their challenges and then report the following task to be *easy* or *medium*. Understanding 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 could be optimised for the learners. It can also help in designing and sequencing the tasks, for the development of effective teaching strategies that can maximize students' learning [42] and reduce undesirable behaviours such as gaming the system [2] and disengagement [29].

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 also thank Dr. Linda Corrin, Donia Malekian and Anam Khan for the useful discussion on this work. 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.

References

1. Arroyo, I., Woolf, B.P., Cooper, D.G. et al.: 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, pp. 506–510 (2011).
2. Baker, R., Corbett, A.T., Koedinger, K.R. et al.: Off-task behavior in the cognitive tutor classroom: when students "game The system". In: SIGCHI Conference on Human Factors in Computing Systems. ACM, pp. 383–390 (2004).
3. Baker, R., D'Mello, S., Rodrigo, Ma.M.T. et al.: 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(4), 223–241 (2010).
4. Bandura, A.: Self-efficacy: The exercise of control. In:Freeman, New York (1997).
5. Bandura, A.: Social Learning Theory. Prentice Hall, Englewood Cliffs, NJ (1977).
6. Belmont, J.M., Mitchell, D.W.: The general strategy hypothesis as applied to cognitive theory in mental retardation. Inrelligence 11(1), 91–105 (1987).
7. Bjork, R.A.: Desirable difficulties perspective on learning. Encyclopedia of the Mind 4, 134–146 (2013).
8. Campbell, D.J.: Subtraction by addition. Memory & Cognition 36(6), 1094–1102 (2008).
9. Campbell, D.J.: Task complexity: A review and analysis. Academy of Management Review 13(1), 40–52 (1988).
10. Carpentar, V.L., Friar, S., Lipe, M.G.: Evidence on the performance of accounting students: Race, gender and expectations. Issues in Accounting Education 8(1), 1–17 (1993).
11. Christensen, T.E., Fogarty, T.J., Wallace, W.A.: The association between the directional accuracy of self-efficacy and accounting course performance. Issues in Accounting Education 17(1), 1–26 (2002).
12. Coştu, B., Ayas, A., Niaz, M.: Investigating the effectiveness of a POE-based teaching activity on students' understanding of condensation. Instructional Science 40(1), 47–67 (2012).
13. Craig, S., Graesser, A., Sullins, J. et al.: Affect and learning: an exploratory look into the role of affect in learning with AutoTutor. Journal of Educational Media 29(3), 241–250 (2004).
14. Csikszentmihalyi, M.: Beyond boredom and anxiety. Jossey-Bass (2000).
15. Csikszentmihalyi, M.: Finding flow: The psychology of engagement with everyday life (1997).
16. Csikszentmihalyi, M.: The flow experience. Consciousness: Brain and states of awareness and mysticism, p. 63–67 (1979).
17. Csikszentmihalyi, M.: Flow: The psychology of optimal experience. Harper Perennial, New York (1990).
18. D'Mello, S., Graesser, A.: Confusion and its dynamics during device comprehension with breakdown scenarios. Acta psychologica 151, 106–116 (2014).

19. D'Mello, S., Graesser, A.: Modeling cognitive-affective dynamics with Hidden Markov Models. In: Annual meeting of the Cognitive Science Society. pp. 2721–2726 (2010).
20. D'Mello, S., Person, N., Lehman, B.: Antecedent-consequent relationships and cyclical patterns between affective states and problem solving outcomes. In: Artifical Intelligence in Education (AIED). pp. 57–64 (2009).
21. D'Mello, S., Taylor, R.S., Graesser, A.: Monitoring affective trajectories during complex learning. In: Annual Meeting of the Cognitive Science Society. pp. 203–208 (2007).
22. D'Mello, S., Graesser, A.: Dynamics of affective states during complex learning. Learning and Instruction 22(2), 145–157 (2012).
23. Dalziel, J.: Practical eTeaching strategies for predict – observe – explain, problem-based learning and role plays. LAMS International, Sydney, Australia (2010).
24. Deloache, J.S., Cassidy, D.J., Brown, A.L.: Precursors of mnemonic strategies in very young children's memory. Child Development 56(1), 125–137 (1985).
25. Dowell, N.M.M., Graesser, A.: Modeling learners' cognitive, affective, and social processes through language and discourse. Journal of Learning Analytics 1(3), 183–186 (2014).
26. Driver, R.: The pupil as scientist? Open University Press, UK (1983).
27. Eccles, J.S., Adler, T.F., Futterman, R. et al.: Expectancies, values and academic behaviors. In: Spence J.T. (ed) Achievement and achievement motives. W.H. Freeman, San Francisco, pp. 75–146 (1983).
28. Eccles, J.S., Wigfield, A.: Motivational beliefs, values, and goals. Annual Review of Psychology 53, 109–132 (2002).
29. Gobert, J.D., Baker, R., Wixon, M.B.: Operationalizing and detecting disengagement within online science microworlds. Educational Psychologist 50(1), 43–57 (2015).
30. Guadagnoli, M.A., Lee, T.D.: Challenge point: A framework for conceptualizing the effects of various practice conditions in motor learning. Journal of Motor Behavior 36(2), 212–224 (2004).
31. Gunstone, R., White, R.: A matter of gravity. Research in Science Education 10, 35–44 (1980).
32. Hom, H.L., Maxwell, F.R.: The impact of task difficulty expectations on intrinsic motivation. Motivation and Emotion 7, 19–24 (1983).
33. Horodyskyj, L.B., Mead, C., Belinson, Z. et al.: Habitable Worlds: Delivering on the Promises of Online Education. Astrobiology 18(1), 86–99 (2018).
34. Kapur, M., Bielaczyc, K.: Designing for productive failure. The Journal of the Learning Sciences 21(1), 45–83 (2012).
35. Kapur, M., Rummel, N.: Productive failure in learning and problem solving. Instructional Science 40(4), 645–650 (2012).
36. Karumbaiah, S., Andres, J.Ma.L., Botelho, A.F. et al.: The implications of a subtle difference in the calculation of affect dynamics. In: International Conference for Computers in Education (2018).
37. Karumbaiah, S., Baker, R., Ocumpaugh, J.: The case of self-transitions in affective dynamics. In: Artificial Intelligence in Education (AIED). pp. 172–181 (2019).
38. Kennedy, G., Lodge, J.M.: 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, pp. 318–328 (2016).
39. Kibirige, I., Osodo, J., Tlala, K.M.: The effect of Predict-Observe-Explain strategy on learners' misconceptions about dissolved salts. Mediterranean Journal of Social Sciences 5(4), 300–310 (2014).

40. Kulhavy, R.W.: Feedback in written instruction. Review of educational research 47(2), 211–232 (1977).
41. Kulhavy, R.W., Yekovich, F.R., Dyer, J.W.: Feedback and response confidence. Journal of Educational Psychology 68(5), 522–528 (1976).
42. Li, W., Lee, A., Solmon, M.: 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(3), 301–318 (2007).
43. Long, M.H., Crookes, G.: Three approaches to task-based syllabus design. TESOL quarterly 26(1), 27–56 (1992).
44. Mangos, P.M., Steele-Johnson, D.: The role of subjective task complexity in goal orientation, self-Efficacy, and performance relations. Human Performance 14(2), 169–185 (2001).
45. Maynard, D.C., Hakel, M.D.: Effects of Objective and Subjective Task Complexity on Performance. Human Performance 10(4), 303–330 (1997).
46. Mooi, T.L.: Self-efficacy and student performance in an accounting course. Journal of Financial Reporting and Accounting 4(1), 129–146 (2006).
47. Nawaz, S., Kennedy, G., Bailey, J. et al.: Moments of confusion in simulation-based learning environments. Journal of Learning Anlytics (in review), (2020).
48. Nawaz, S., Kennedy, G., Bailey, J. et al.: 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, pp. 224–233 (2018).
49. Pintrich, P.R., Schunk, D.H.: Motivation in education: Theory, research, and applications. Prentice Hall (2002).
50. Robinson, P.: 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), 1–33 (2005).
51. Robinson, P.: 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, pp. 185–316 (2001).
52. Robinson, P.: Task complexity, task difficulty, and task production: Exploring interactions in a componential framework. Applied Linguistics 22(1), 27–57 (2001).
53. Rodrigo, Ma.M.T., Baker, R., Agapito, J. et al.: The effects of an interactive software agent on student affective dynamics while using an intelligent tutoring system. In: IEEE Transactions on Affective Computing. pp. 224–236 (2012).
54. Schneider, D.W., Anderson, J.R.: Asymmetric switch costs as sequential difficulty effects. The Quarterly Journal of Experimental Psychology 63(10), 1873–1894 (2010).
55. Shute, V.J., D'Mello, S., Baker, R. et al.: Modeling how incoming knowledge, persistence, affective states, and in-game progress influence student learning from an educational game. Computers & Education 86, 224–235 (2015).
56. Skehan, P. (1998) A cognitive approach to language learning. Oxford University Press
57. Sreerekha, S., Arun, R.R., Sankar, S.: Effect of Predict-Observe-Explain strategy on achievement in chemistry of secondary school students. International Journal of Education & Teaching Analytics 1(1), 1–5 (2016).
58. Stephanou, G., Kariotoglou, P., Dinas, K.D.: 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(1), 45–72 (2011).

59. Tao, P.K., Gunstone, R.F.: The process of coneptual change in force and motion during computer-supported physics instruction. Journal of Research in Science Teaching 36(7), 859–882 (1999).
60. Vygotsky, L.S.: Mind and society: The development of higher mental processes. Harvard University Press, Cambridge, MA (1978).
61. White, R., Gunstone, R.: Probing understanding. Routledge (1992).
62. Zou, X., Ma, W., Ma, Z. et al.: Towards Helping Teachers Select Optimal Content for Students. In: International Conference on Artificial Intelligence in Education (AIED). Springer, Chicago, IL, pp. 413–417 (2019).