Struggle town? Developing profiles of student confusion in simulation-based learning environments

Sadia Nawaz
University of Melbourne
Australia

Gregor Kennedy
University of Melbourne
Australia

James Bailey
University of Melbourne
Australia

Chris Mead
Arizona State University
USA

Lev Horodyskyj
Arizona State University
USA

A considerable amount of research on emotions and learning has been undertaken in recent years. Confusion has been noted as a particularly important emotion as it has the potential to trigger students' engagement in learning tasks. However, unresolved confusion may turn into frustration, boredom and ultimately disengagement. The study reported in this paper investigated whether learning analytics could be used to successfully determine indicators or patterns of interactions that may be associated with confusion in a simulation-based learning environment. The findings of the study indicated that when taken individually, measures on specific learning tasks only hint at when students are struggling, but when taken together these indicators present a pattern of student interactions or a student profile that could be indicative of confusion.

Keywords: simulation, learning analytics, confusion, predict-observe-explain, learning process

Introduction

Digital learning environments (DLE) are becoming pervasive in higher and tertiary education as they can offer scalable, economical educational activities for both teachers and students. While on the one hand simulation-based environments, depending on their design, can present students with exploratory and relatively unstructured learning experiences, there is a significant chance for students to become confused due to the absence of immediate guidance and feedback, either from the teachers or by the system (Pachman, Arguel, & Lockyer, 2015). Confusion is an epistemic emotion (Pekrun, 2010; Pekrun, Goetz, Titz, & Perry, 2002) – an emotion which arises when learning is taking place. Other epistemic emotions that may arise during the learning process include, surprise, delight, curiosity, as well as anxiety, frustration and boredom (Baker, D'Mello, Rodrigo, & Graesser, 2010; Calvo & D'Mello, 2010; D'Mello & Graesser, 2012). Understanding how students experience these emotions in DLEs is increasingly important for enhancing the design of these environments.

Prior research has shown that emotions play an important role in learning, motivation, development and memory (Ainley, Corrigan, & Richardson, 2005; Ashby, Isen, & Turken, 1999; Isen, 1999; Lewis & Haviland-Jones, 2004). Confusion is particularly important as it can arise in complex learning tasks that require students to make inferences, solve advanced problems, and demonstrate application and transfer of knowledge. Research has shown that in complex learning activities, confusion is 'unlikely to be avoided' (D'Mello, Lehman, Pekrun, & Graesser, 2014) and the resolution of confusion requires students to stop, think, reflect and review their misconceptions (D'Mello & Graesser, 2012). While confusion can be beneficial to learning, unresolved or prolonged confusion may leave a student feeling stuck and frustrated (Baker et al., 2010; Calvo & D'Mello, 2010). Such frustration can ultimately transition into boredom which can lead to students disengaging from the task (D'Mello & Graesser, 2012), a critical point which educators aim to prevent (D'Mello & Graesser, 2014b; Liu, Pataranutaporn, Ocumpaugh, & Baker, 2013). Thus, sustained unresolved confusion is detrimental to learning and has been associated with negative emotional oscillations (D'Mello & Graesser, 2014a; D'Mello & Graesser, 2012; D'Mello et al., 2014). D'Mello and Graesser dubbed the balance between creating 'useful' confusion for students and not making them too confused the 'zone of optimal confusion' (D'Mello & Graesser, 2014a).

While persistent confusion needs to be avoided, some learning designs aim to promote a degree of difficulty that is likely to result in confusion. These include teaching and learning frameworks such as problem-based learning

(Schmidt, 1983), device breakdown (D'Mello & Graesser, 2014b) and productive failure (Kapur, 2016). Another common learning design which can inherently promote confusion is the simulation-based, predict-observe-explain (POE) paradigm (White & Gunstone, 1992). POE is a three-sequence design where: (i) during the prediction phase students develop a hypothesis about a conceptual phenomenon, and state their reasons for supporting that hypothesis (ii) during the observe phase students explore an environment related to the conceptual phenomenon, view data, and see what 'actually' happens and finally, (iii) during the explain phase the ideas and concepts related to the phenomenon are explained and elaborated, and the reasoning about the conceptual phenomenon is provided to the students. It is likely that students in a POE environment may feel confused, particularly when there is a discrepancy between their current understanding (predictions) and what they find out (observations) while completing a simulation.

POE environments have mostly been used to investigate students' prior knowledge and misconception (Liew & Treagust, 1995) as well as to investigate the effectiveness of these environments in terms of peer learning opportunities (Kearney, 2004; Kearney, Treagust, Yeo, & Zadnik, 2001) and conceptual change (Tao & Gunstone, 1999). In our recent work (Kennedy & Lodge, 2016), a simulation-based environment was used to study students' self-reported emotional transitions. This study found that a POE based environment could help students overcome their initial misconceptions through feedback and scaffolding. The current study adds to this research by investigating whether learning analytics-based markers can be used to detect patterns of interactions that might suggest students are "struggling" or confused in a simulation-based POE environment.

The use of analytics in DLEs have been used for some time to investigate students' learning processes but have risen in prominence lately (Campbell, DeBlois, & Oblinger, 2007; Goldstein & Katz, 2005; Kennedy, 2004; Kennedy, Ioannou, Zhou, Bailey, & O'Leary, 2013; Kennedy & Judd, 2004, 2007) . The use of analytics to understand emotions in DLEs has received less attention in the literature (Lee, Rodrigo, d Baker, Sugay, & Coronel, 2011; Liu et al., 2013). Measuring or detecting emotions such as confusion is inherently difficult because, as an emotion, confusion can be relatively short-lived (D'Mello & Graesser, 2014a), unlike some of the emotions which sustain over a longer period (e.g. boredom; see (D'Mello et al., 2014)). Detecting confusion in naturalistic learning environments is also challenging as these environments restrict the way data can be collected, particularly in comparison to lab-based environments where sensors, physiological trackers, emote-aloud protocols, video recordings and many other data collection tools and techniques can be used (D'Mello & Graesser, 2014a). Moreover, relying solely on self-report measures of confusion can be 'insensitive' (D'Mello et al., 2014) and problematic due to 'intentional' misreporting (Komar, Brown, Komar, & Robie, 2008; Tett, Freund, Christiansen, Fox, & Coaster, 2012) which the students might do to avoid social pressure (Kennedy & Lodge, 2016). Therefore, the aim of this study was to investigate whether learning analytics could be successfully used to determine indicators of or patterns of interactions that may be associated with confusion in a POE, simulation-based learning environment.

Habitable Worlds

The DLE used in this research is called *Habitable Worlds* – an introductory science class that covers foundational concepts in biology, physics and chemistry (Horodyskyj et al., 2018). *Habitable Worlds* is a project-based **course** that encourages students to solve problems using logic and reasoning and promotes students' engagement using interactive tasks. The course is built using Smart Sparrow – an adaptive eLearning platform, which makes it possible to track students' learning activities and interactions. *Habitable Worlds* consists of 67 interactive **modules**, several of which are based on the POE protocol. *Stellar Lifecycles* is one of the first POE modules in the course and it was a primary focus in this study. In this module, several **tasks** were embedded that spanned 23 screens. A task in this context refers to a number of activities students are asked to complete on any given screen. These activities may include free-text answers to a question, watching videos, completion of a multiple-choice questions, or the "submissions" associated with interacting with simulations. For this paper, students learning interactions at the module and task level were analysed.

Students were asked to engage in a series of learning activities, the primary sequence of which is provided below.

- View an explanatory video about different objects in our universe and how they differ in sizes.
- Students then need to select a hypothesis about what they think the relationship between stellar lifespan and stellar mass is from five possible choices (i.e. make a prediction) and also report through free-text their reasons for selecting their hypothesis. Notably, students are not provided with any content relating to this question prior to this.
- Students next use a simulator to explore, and hopefully develop an understanding of, the relationship between stellar lifespan with stellar mass. Students use the simulator to create and manipulate virtual

stars, so they can observe the mass and the relative lifespans of stars. They can use the simulator as many times as they wish and each "run" of the simulation is recorded as a submission.
- After becoming familiar with the simulator, students are asked to engage with two more complex tasks: creating virtual stars of a given mass range and reporting on the lifespan of these stars. Again, students can use the simulator as many times as they wish and each run of the simulation is recorded as submission. After completing the simulation and associated questions students are then prompted to either accept or reject their earlier proposed hypothesis.
- The follow up task, which is only available to those students who had predicted an incorrect hypothesis and endorsed this prediction, asks students to update their hypotheses. Students cannot complete this screen without selecting the correct hypothesis; in effect the program narrows all options until the student chooses the correct one.
- Towards the end of the sequence of activities students are asked to watch a video that provides them with a complete explanation of the relationship between stellar lifespan and mass. On this screen each student's first proposed hypothesis is reproduced, as is the correct hypothesis and estimates of stellar lifespans for the various star classes.
- The final set of screens asks students to create and burn different virtual stars. These tasks require students to make observations on the *HertzSprung-Russell* diagram, which shows the changes in a star's colour, luminosity, temperature and classification. Students are asked to make decisions and selections about the stages through which stars go as they age.

It is important to note that the program was "adaptive"; which in this context generally meant the program provided students with feedback and hints on their responses (or lack of response). It also typically meant that students were not allowed to progress or move on until a task had successfully been completed.

Methodology

A total of 364 science undergraduate students from a large US-based university attempted *Stellar Lifecycles* as part of their undergraduate study. Over 15,000 interaction entries were recorded within the digital learning environment and these interactions formed the basis of the data collected for study. A range of measures were used, based on analytics recorded from the system, to develop patterns of interaction with the system. The measures used in the analyses are presented in detail in the results section but included measures such as time on task, attempts at tasks, accuracy of attempts at tasks, and content analyses of free-text responses. The analysis presented in the Results section used an iterative analytics approach consistent with that proposed by Kennedy and Judd (Kennedy & Judd, 2004).

Results and discussion

Module level patterns

Data analysis began with pre-processing and outlier elimination, which involved removing all individual measures that were outside five standard deviations from the median. An initial cluster analysis was undertaken at the *Stellar Lifecycles* ***Module*** level to determine students' general engagement patterns. Variables included in this cluster analysis were mean module score, mean module completions, mean attempts on module tasks, and mean time on module. A three-cluster solution was the clearest description of the data. However, it was clear that the third cluster, which contained only 22 students, were those students who had very low mean module scores, task attempts across the module, and mean time on the module. These students did not complete the module – they exited the module at the halfway point – and as a result they were removed from further analyses. The profiles of the remaining two clusters are presented in Table 1.

Table 1: Learners' overall engagement patterns in Stellar Lifecycles.

	Cluster 1 (n =212)		Cluster 2 (n=130)		T	p
	Mean	SD	Mean	SD		
Module scores	11.99	0.15	12	0	-1	0.32
Module task completions	0.9	0.04	0.8	0.04	20.03	<.001
Attempts at module tasks	16.7	3.8	29.4	8.83	-15.99	<.001
Time on module (mins)	140.13	163.56	258.63	644.61	-2.05	0.04

Table 1 shows that both clusters of students achieved the maximum score for the module, completing all the required tasks. However, students in Cluster 1 had a significantly higher number of task completions compared to students in Cluster 2 and students in Cluster 2 had significantly more task attempts and spent longer on module tasks. So, while students in both clusters were achieving the same end, they seemed to follow different processes getting there. This high-level data could be interpreted in many ways. One could be that students in Cluster 2 could be diligent and dedicated students who spent more time and had more attempts at tasks in the module, leading to success that was commensurate with those from Cluster 1 who seemed, for whatever reason, to arrive at the same end point "more easily". Alternatively, students in Cluster 2 could have struggled more and have been more confused about their engagement with the module and its content compared to those in Cluster 1; and this struggle and confusion was manifest in their behavioural data, notably more attempts at tasks and taking longer to complete tasks.

We used this second working hypothesis to frame subsequent analyses. That is, we were keen to see whether other learning analytics-based markers at both the module and the task or screen level could help to further discriminate and characterise the two groups of students that had emerged from the cluster analysis at the module level.

Response time to module tasks

The next set of analyses concentrated on the average time students were taking to complete tasks presented to them across the module. To undertake this analysis students' responses to tasks across all the screens of *Stellar Lifecycles* were analysed. The mean time students took to make each task attempt is presented in Figure 1. Figure 1 shows the number of attempts on the X axis (some students made as many as 10 attempts) and the mean time taken to make each attempt on the Y axis. It can be seen that students in both Cluster 1 and Cluster 2 were, on average, slower when making their first attempt at a task compared to their subsequent responses. It is also clear that Cluster 1 students were initially responding more quickly to tasks than Cluster 2 students. There may be a number of reasons for this: students in Cluster 1 may be more confident, and/or have higher prior knowledge than those in Cluster 2; conversely students in Cluster 2 may be more careful and/or more unsure or more confused about their response to the tasks. What is also noticeable from Figure 1 is that there is a general reduction in response time across attempts for Cluster 1 students, and there are clear spikes of response time for Cluster 2 students (attempt 4 and attempt 7). This may also be indicative of Cluster 2 students being more uncertain or confused about what their response to the task should be.

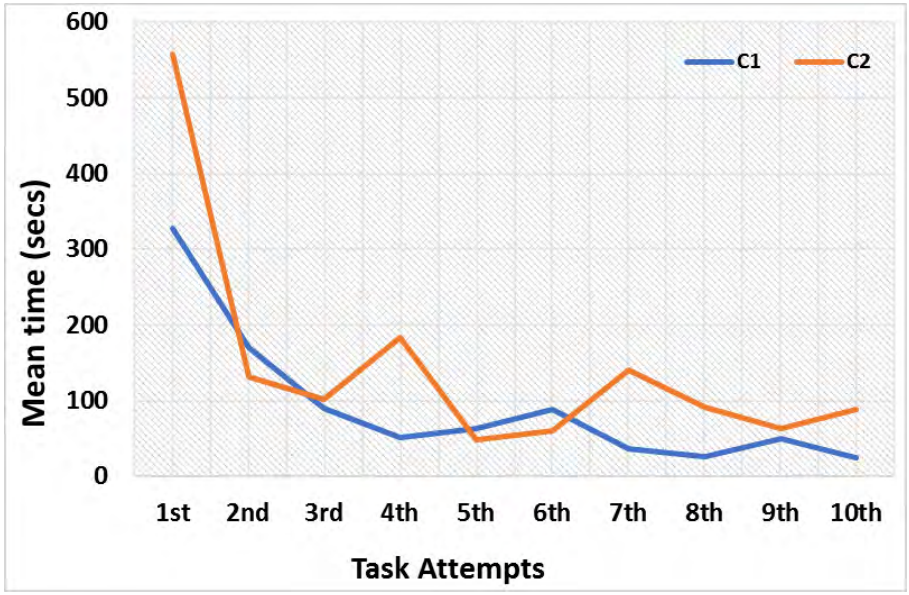

Figure 1: Analysis of mean response time per task attempt.

Task level patterns

The next set of analyses considered students' interactions at the screen or task level rather than the module level.

Students' initial predictions

The next set of analyses examined the nature of hypothesis being selected by students in the *Predict* phase of the module. Overall, it can be seen that there were no differences between the two clusters on their hypothesis

selections. Approximately two-thirds of students, regardless of cluster, chose an incorrect hypothesis. We anticipated that many students would have a common misconception about the relationship between the size of a star and its lifespan (i.e. they would intuitively believe bigger stars live longer). The results from Table 2 indicate this to be the case with large numbers of students in both clusters endorsing this lure (Cluster 1 = 42%, Cluster 2 = 49%). These results suggest that groups had similar levels of prior knowledge before beginning the simulation-based POE task, and many held a common misconception.

Table 2: Students' hypotheses during the *Prediction* phase (by cluster).

	Cluster 1	Cluster 2		
	N (%)	N (%)	T	p
Correct hypothesis	72 (34.0%)	41(31.5%)	0.46	0.65
Incorrect hypothesis	137 (64.6%)	87 (66.9%)	-0.46	0.65
Endorsing common misconception hypothesis	89 (42.0%)	64 (49.2%)	-1.32	0.19
Endorsing other misconception hypotheses	48 (22.9%)	23 (18.0%)	1.11	0.27

Students' detailed prediction behaviours

Next, a more detailed set of analyses considered students' responses to the prediction phase of the module. It can be seen from Table 3, that on average Cluster 1 students were spending a little over two and a half minutes on this screen, while Cluster 2 students were spending on average over 11 minutes. While not statistically different (most likely due to the high standard deviation for Cluster 2 students) this seems to represent a clear qualitative difference between the two groups. It can also be seen from Table 3 that when students were making their prediction about the relationship between stellar lifespan and stellar mass, students in Cluster 2 were making significantly more attempts at the hypothesis selection than students in Cluster 1. The most common reason was that the length of students' text response used to justify their hypothesis was too short and they were asked to resubmit it. This interpretation is consistent with number of words written overall, as the mean word count for Cluster 2 students was significantly lower than Cluster 1 students. Finally, when both the clusters were compared in terms of unique words per person, Cluster 1 students used more unique words per person.

Table 3: Students' engagement patterns during the *Prediction* phase (by cluster).

	Cluster 1		Cluster 2			
	Mean	SD	Mean	SD	T	p
Time (secs)	141.75	277.52	669.89	6114.57	-1.04	0.30
Attempts	1.06	0.25	1.13	0.36	-1.94	0.05
Word count	16.19	11.15	12.77	10.03	-3.78	0.001
Unique words (per student)	14.35	9.00	11.5	7.48	3.14	0.001

As described above, after students had made an initial hypothesis selection they were asked to justify why they believed their hypotheses to be true in a free-text response. Stop word elimination (i.e. eliminating words like "a" "it" 'the" "is") was completed for all student text responses and the primary keywords for each cluster were then determined through a content analysis. In order to compare the rank and relative frequency of keywords across clusters, the percentage of times these words appeared in each cluster were calculated. **Error! Reference source not found.** presents a bubble plot where the words are arranged in descending rank order of frequency. Such frequency-based analyses have been used in various other disciplines (Nawaz & Strobel, 2016; Nawaz, Usman, & Strobel, 2013).

Content analysis of students' justification of their prediction

Error! Reference source not found. shows clear similarity between the words used by Cluster 1 and 2 students to justify their hypotheses. For example, words such as "*star*", "*mass*", "*energy*" and "*long*" are the highest ranked and most frequently used words by students in both clusters. Beyond this, there are some differences between students in each cluster. While it is important not to overstate these differences, they are useful to note as a profile of students' interaction and engagement is established across the module.

First, while both Cluster 1 and Cluster 2 students use the word "*guess**", it is used more slightly more frequently by Cluster 2 students (1.38% of all words) than for Cluster 1 (0.84%). We assume the presence of the word "*guess*" means that some of the students in these clusters were unsure of their hypothesis or perhaps they were uncertain of why that hypothesis holds. In support of this conclusion the word "*guess**" consistently co-occurs with "*just*" and an analysis of raw text commonly revealed phrases such as "it's just a guess", "seems to make sense but it was just a guess", and "not sure, this is just a guess". The analysis also considered the occurrence of technical terms in students' responses so that a judgement could be made about the quality or clarity of students responses (DeGroff, 1987). While Cluster 1 students used a number of key terms (such as "*fusion*", "*fused*", "*hot*", "*sustain*", "*bright*") these words were largely absent from the word profiles of Cluster 2 students. Interestingly, these words also appeared in the lecture material explaining the relation between stellar mass and its lifespan. The use of such terms by Cluster 1 students could be indicative of more understanding or content awareness of these students.

The content of students' text-based responses suggests that while the most common words are similar across clusters, there tend to be differences thereafter. Compared to students from Cluster 1, Cluster 2 students tend use the word "guess" slightly more and tend to use technical terms slightly less.

Observation of events and change in hypothesis

The *Observe* phase provided a basic introduction to the stellar simulator and guided students on how to create and run stars of varying solar mass. The second part of this phase required students to create stars of specific mass range and then report on the associated lifespans. While several students found this difficult – entering their observed data – the adaptive feedback ensured that all students eventually entered the data correctly. The interaction patterns for this task, recorded via analytics, showed that students in Cluster 2 spent significantly less time on this task (Cluster 2: M =148.52 (149.10); Cluster 1: M = 252.34 (477.11); $T(309) = 3.21$; $p < .001$) but made more attempts at the task before completing it (Cluster 2: M =1.58 (1.18); Cluster 1: M = 1.34 (1.26); $T(406) = -2.04$; $p < .05$). This is difficult to interpret with confidence but could suggest that Cluster 1 students took a more considered approach to this task, particularly given students in Cluster 2 completed the task more quickly with more errors (which may be indicative of rapid trial and error behaviour).

Once the values were correctly recorded, students were then asked to report whether they would like to accept or reject their earlier proposed hypotheses. **Error! Reference source not found.** shows the percentage of students in each cluster who maintained or rejected their initial correct or incorrect hypotheses. While Cluster 1 students were more likely to maintain a correct hypothesis and to reject an incorrect hypothesis, there was a small fraction in both clusters who did not respond to this question on their first attempts.

While **Error! Reference source not found.** showed the percentage of students who rejected their incorrect hypotheses, it will also be useful to consider the new hypotheses proposed by these students and whether students subsequently proposed a correct hypothesis. We found that all students in Cluster 1 who had first proposed an incorrect hypothesis revised this so that it was subsequently correct. A large proportion of Cluster 2 students also did this, but it is worth noting that despite the program effectively directing them – using adaptive feedback – to the relationship between stellar size and lifespans, six students from the Second Cluster (6.7% of those who proposed an incorrect hypothesis) revised their hypothesis so that it still was incorrect.

Mean explanation errors

Toward the end of the module students were provided with an explanation of the concepts they were learning about in the module. Part of this section of the module asked students to complete a task that would demonstrate their understanding of the minimum and maximum lifespans of seven different classes or types of stars. In completing the task, a total of 14 different values needed to be entered and students could submit responses as many times as they wanted. Each time a response set was submitted students received adaptive feedback which guided them and helped them complete the task. If students did not enter any values this would result in the maximum number of errors for the task being recorded (reflected in a score of 7).

As students spent more time with the task and entered more responses, the number of errors would diminish (i.e. students would change their incorrect responses). It was expected that after a series of attempts students would gradually reduce their number of errors so that eventually there would be no incorrect responses.

The mean explanation errors for students in Cluster 1 and Cluster 2, at successive task attempts, are presented in Figure 2. Students in both clusters gradually reduced their errors over time. It can also be seen that students in Cluster 1 started with fewer errors than the students in Cluster 2. Moreover, it is clear that students in Cluster 1 reached a resolution to the task in fewer attempts than students in Cluster 2.

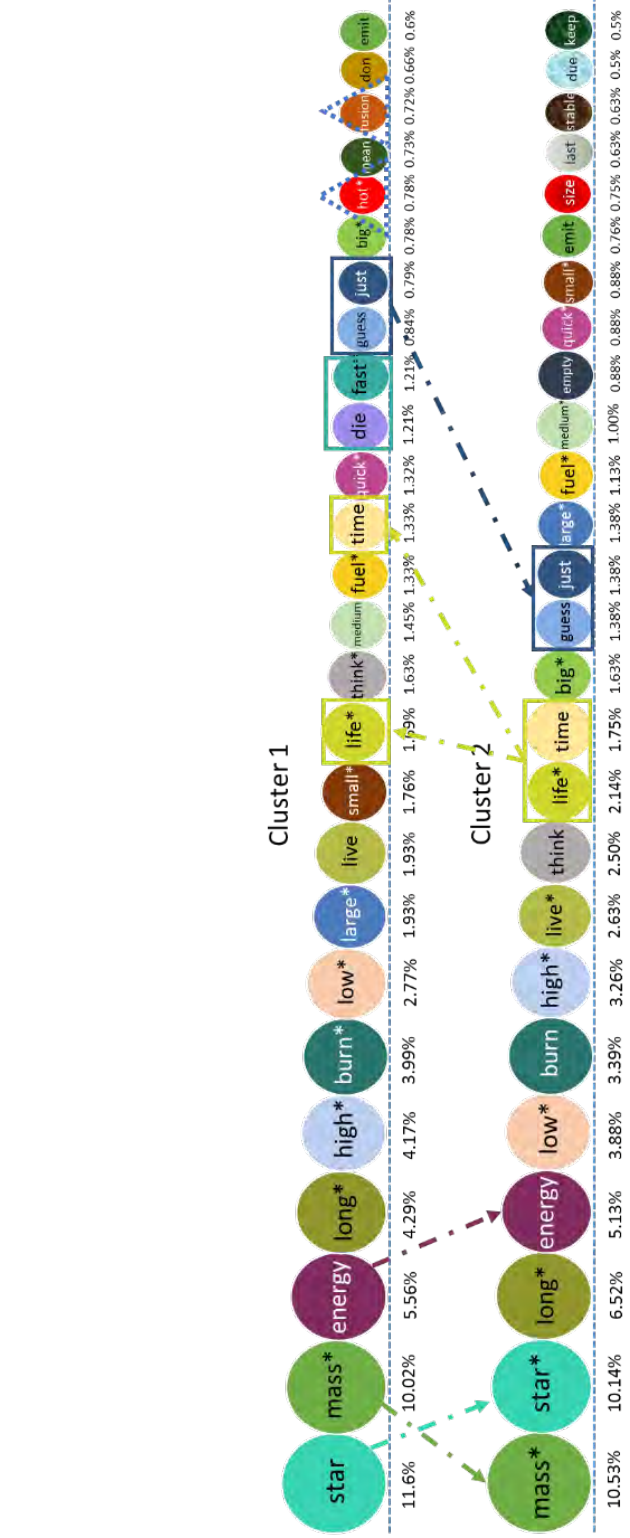

Figure 2: Keyword analysis for assessing students' text response for hypothesis justification.

Table 4: Maintenance and rejection of students' initial hypotheses during the *Observation* phase (by cluster).

	Cluster 1	Cluster 2
Initial correct hypothesis *maintained* (1st attempt)	70 (97.2%)	36 (87.8%)
Initial incorrect hypothesis *maintained* (1st attempt)	1 (0.7%)	8 (9.2%)
Initial correct hypothesis *rejected* (1st attempt)	2 (2.8%)	2 (4.9%)
Initial incorrect hypothesis *rejected* (1st attempts)	134 (97.8%)	78 (89.6%)
Initial correct hypothesis *untested* (1st attempts)	0 (0.0%)	3 (7.3%)
Initial incorrect hypothesis *untested* (1st attempts)	2 (1.5%)	1 (1.2%)

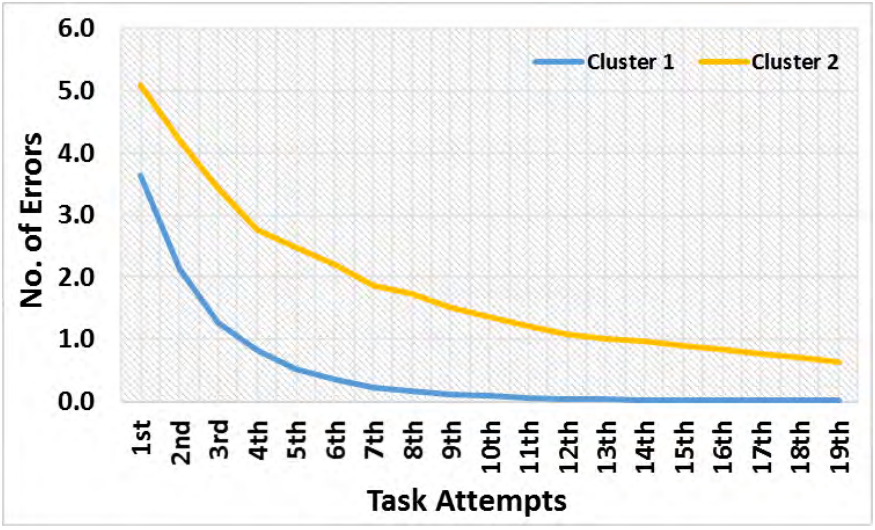

Figure 2: The number of student errors in the explanation task by task attempts

A final task in this section of the module asked students to make careful observations of how, when a star dies, it changes in luminosity, temperature and stellar classification. The tasks spanned three screen and on each screen the students needed to create and run stars of different stellar classes (types) and mass. For example, students might be asked to create a "red dwarf" with a solar mass between 0.08 and 0.49. Students were asked to indicate which of the four stellar class(es) the star went through as it aged (i.e. Giant Star, Super-Giant, White Dwarf and Supernova). The data from students' interactions indicated that on 103 occasions, no response was provided by the students on submission. When analysed by cluster, it was clear that students in Cluster 2 were significantly more likely not to provide a response to this final activity compared to students in Cluster 1 (Cluster 2: 22.3%; Cluster 1: 7.1%).

Students' conceptual understanding

The final set of analyses considered whether there were differences between clusters of students when it came to their conceptual understanding of the content of the module. While students' initial hypotheses suggest that the two student clusters came into the module with more or less similar understanding (and misconceptions) about the relationship between star size and lifespan, we were keen to assess students' understanding at the end of the module. Conceptual understanding was assessed using a complex transfer task that was presented to students in a separate module of *Stellar Lifecyles* called *Stellar Applications*. In this task, students were asked to calculate the properties of six stars (properties such as luminosity, temperature, and mass) and identify the longest-lived and shortest-lived star. A total of 10 points were available for completely correct answers and students could complete the task multiple times but were penalised for incorrect attempts. A T-test that compared students' scores on this measure of conceptual understanding indicated that students in Cluster 1 (M=7.54; SD = 3.18) showed greater understanding than students in Cluster 2 (M = 6.61; SD = 3.67) ($T(310) = 2.35$; $p < .01$).

Conclusion

The first empirical finding presented in this paper was that a module-level cluster analysis revealed two distinct groups of students who, broadly speaking, completed a simulation-based learning task in different ways. While both groups were successful – in part because the adaptive nature of the program ensured it – the learning process they went through to achieve this success seemed to differ on generalised metrics. Further analyses of students' completion of all tasks in the module and their screen-based interactions showed a number of other differences between clusters. When viewed discretely many of these screen-based differences were only modest. However, when viewed collectively or in aggregate, these discrete screen-based differences revealed patterns of interaction that allow students from the two clusters to be distinguished and potentially characterised.

Overall, Cluster 1 students tended to respond to tasks more quickly, arrive at their hypothesis more quickly and tended to write more and more technically about it. Students in Cluster 1 spent more time observing the data from the simulation and made less errors in their observations. In contrast, students in Cluster 2 tended to take more time to respond to tasks, took more time to arrive at a hypothesis, and when they did, they seemed more unsure of it. They spent less time observing the outcomes of the simulation presented and they made more errors

than those students in Cluster 1. While many students in both clusters rejected an initial incorrect hypothesis, students in Cluster 2 seemed less likely to do this. When it came to the final explanation of the phenomenon, students in Cluster 1 made fewer errors from the start and corrected their errors more quickly. Those in Cluster 2 started with significantly more errors and took longer to correct them, even with the adaptive feedback and support provided by the program. Finally, students in Cluster 1 understood the material covered significantly more than those in Cluster 2. It seems unlikely that these differences in the learning interactions and learning process can be attributed to students in Cluster 1 having greater prior knowledge as both groups made similarly poor predictions at the start of the task and had similar levels of misconception.

What the patterns of interactions do suggest is that students in Cluster 2, for whatever reason, struggled with what was being asked of them in the module. They seemed to find learning more difficult as a process – as measured by analytics markers of their various interactions with different tasks – and this was reflected in their learning outcome. These signs of "struggle" could also be interpreted as signs of confusion. While the pattern of interactions observed for Cluster 2 students – taking a long time to respond to tasks, not being able to quickly correct errors, finding it hard to explain responses – could be attributed to disengagement, we contend that this pattern could as easily be consistent with the profile of a student who is confused and struggling with the learning content and task. But it is, of course, not possible to be definitive about this, based on a single study.

The next steps in this program of research will be to consider the ways in which learning analytics may be used to generate markers of specific moments of student confusion in simulation-based POE environments. That is, it is likely that students would experience confusion when they realise there is a mismatch between their initial prediction or hypothesis and what they then observe in a simulation-based environment. The findings about students' general patterns of interactions presented in this paper – indicating that some students are struggling while others struggle less – provide an excellent context for these more detailed analyses.

Acknowledgements

This work was partially supported by the Australian Government Research Training Scholarship (RTS) and the Science of Learning Research Scholarship (SLRC). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the funding agencies. We are thankful to the colleagues Jason and Paula at the Melbourne Centre for the Study of Higher Education for their valuable comments and feedback on this research.

References

Ainley, M., Corrigan, M., & Richardson, N. (2005). Students, tasks, and emotions: Identifying the contribution of emotions to students' reading of popular culture and popular science texts. *Learning and Instruction, 15*, 433– 447.

Ashby, F. G., Isen, A. M., & Turken, A. U. (1999). A neuropsychological theory of positive affect and its influence on cognition *Psychological Review, 106*, 529-550. https://doi.org/10.1037//0033-295X.106.3.529

Baker, R. S., D'Mello, S., Rodrigo, M. M. T., & Graesser, A. (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*(4), 223-241.

Calvo, R. A., & D'Mello, S. (2010). Affect Detection: An Interdisciplinary Review of Models, Methods and Their Applications. *IEEE Transactions on Affective Computing, 1*(1), 18-37.

Campbell, J. P., DeBlois, P. B., & Oblinger, D. G. (2007). Academic analytics: A new tool for a new era. *EDUCAUSE, 42*(4), 40.

D'Mello, S., & Graesser, A. (2014a). Confusion. In R. Pekrun & L.-g. Lisa (Eds.), *International Handbooks of Emotions in Education* (pp. 289-310). Hoboken: Taylor and Francis.

D'Mello, S., & Graesser, A. (2014b). Confusion and its dynamics during device comprehension with breakdown scenarios. *Acta psychologica, 151*, 106-116. https://doi.org/10.1016/j.actpsy.2014.06.005

D'Mello, S., & Graesser, A. (2012). Dynamics of affective states during complex learning. *Learning and Instruction, 22*(2), 145-157.

D'Mello, S., Lehman, B., Pekrun, R., & Graesser, A. (2014). Confusion can be beneficial for learning. *Learning and Instruction, 29*, 153-170. https://doi.org/10.1016/j.learninstruc.2012.05.003

DeGroff, L.-J. C. (1987). The Influence of Prior Knowledge on Writing, Conferencing, and Revising. *The Elementary School Journal, 88*(2), 105-118. https://doi.org/10.1086/461527

Goldstein, P. J., & Katz, R. N. (2005). Academic analytics: The uses of management information and technology in higher education. *EDUCAUSE, 8*, 1-12.

Horodyskyj, L. B., Mead, C., Belinson, Z., Buxner, S., Semken, S., & Anbar, A. D. (2018). Habitable Worlds: Delivering on the Promises of Online Education. *Astrobiology, 18*(1), 86-99. doi:10.1089/ast.2016.1550

Isen, A. M. (1999). *Positive affect*. New York: Wiley. https://doi.org/10.1002/0470013494.ch25

Kapur, M. (2016). Examining productive failure, productive success, unproductive failure and unproductive success in Learning. *Educational Psychologist, 0*(0), 1-11. doi:http://dx.doi.org/10.1080/00461520.2016.1155457
Kearney, M. (2004). Classroom use of multimedia-supported predict–observe–explain tasks in a social constructivist learning environment. *Research in science education, 34*(4), 427-453. https://doi.org/10.1007/s11165-004-8795-y
Kearney, M., Treagust, D. F., Yeo, S., & Zadnik, M. G. (2001). Student and teacher perceptions of the use of multimedia supported predict–observe–explain tasks to probe understanding. *Research in science education, 31*(4), 589-615.
Kennedy, G. (2004). Promoting cognition in multimedia interactivity research. *Journal of Interactive Learning Research, 15*(1), 43-61.
Kennedy, G., Ioannou, I., Zhou, Y., Bailey, J., & O'Leary, S. (2013). Mining interactions in immersive learning environments for real-time student feedback. *Australasian Journal of Educational Technology, 29*(2).
Kennedy, G., & Judd, T. S. (2004). Making sense of audit trail data. *Australasian Journal of Educational Technology, 20*(1).
Kennedy, G., & Judd, T. S. (2007). Expectations and reality: Evaluating patterns of learning behaviour using audit trails. *Computers & Education, 49*(3), 840-855.
Kennedy, G., & Lodge, J. M. (2016, November). *All roads lead to Rome: Tracking students' affect as they overcome misconceptions.* Paper presented at the ASCILITE, Adelaide. https://doi.org/10.14742/apubs.2016.812
Komar, S., Brown, D. J., Komar, J. A., & Robie, C. (2008). Faking and the validity of conscientiousness: A Monte-Carlo investigation *Journal of Applied Psychology, 93*, 140–154. https://doi.org/10.1037/0021-9010.93.1.140
Lee, D. M. C., Rodrigo, M. M. T., d Baker, R. S., Sugay, J. O., & Coronel, A. (2011, October 2011). *Exploring the relationship between novice programmer confusion and achievement* Paper presented at the In International Conference on Affective Computing and Intelligent Interaction, Berlin, Heidelberg.
Lewis, M., & Haviland-Jones, J. M. (2004). *Handbook of emotions* (Vol. 2). New York: Guilford Press.
Liew, C. W., & Treagust, D. (1995). A Predict-Observe-Explain Teaching Sequence for Learning about Students' Understanding of Heat and Expansion Liquids. *Australian Science Teachers' Journal, 41*(1), 68-71.
Liu, Z., Pataranutaporn, V., Ocumpaugh, J., & Baker, R. (2013, July 2013). *Sequences of frustration and confusion, and learning.* . Paper presented at the Educational Data Mining.
Nawaz, S., & Strobel, J. (2016). Authorship and Content Analysis of Engineering Education Research: A Case Study. *International Journal of Engineering Pedagogy (iJEP), 6*(2), 39-51.
Nawaz, S., Usman, M., & Strobel, J. (2013). Analysis of the influence of the International Journal of Electrical Engineering Education on electrical engineering and electrical engineering education. *International Journal of Electrical Engineering Education, 50*(3), 316-340. https://doi.org/10.3991/ijep.v6i2.5577
Pachman, M., Arguel, A., & Lockyer, L. (2015, November 29- December 2). *Learners' confusion: faulty prior knowledge or a metacognitive monitoring error?* Paper presented at the ACILITE, Perth, Australia.
Pekrun, R. (2010). *Academic emotions* (Vol. 2). Washington, DC: American Psychological Association.
Pekrun, R., Goetz, T., Titz, W., & Perry, R. P. (2002). Academic emotions in students' self-regulated learning and achievement: A program of quantitative and qualitative research. *Educational Psychologist, 37*, 91–106.
Schmidt, H. G. (1983). Problem-based learning: Rationale and description. *Medical education, 17*(1), 11-16.
Tao, P., & Gunstone, R. (1999). The process of conceptual change in force and motion during computer-supported physics instruction. . *Journal of Research in Science Teaching, 36*(7), 859–882.
Tett, R. P., Freund, K. A., Christiansen, N. D., Fox, K. E., & Coaster, J. (2012). Faking on self-report emotional intelligence and personality tests: Effects of faking opportunity, cognitive ability, and job type. *Personality and Individual Differences, 52*, 195–201. doi:10.1016/j.paid.2011.10.017
White, R., & Gunstone, R. (1992). *Probing understanding*: Routledge.

Please cite as: Nawaz, S., Kennedy, G., Bailey, J., Mead, C. & Horodyskyj, L. (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, H. Suri & J. Tai (Eds.), Open Oceans: Learning without borders. Proceedings ASCILITE 2018 Geelong (pp. 224-233). https://doi.org/10.14742/apubs.2018.1905