July, 2014

Can engagement be compared? Measuring academic engagement for comparison

Ling Tan, *ACER*
Xiaoxun Sun, *ACER*
Siek Toon Khoo, *ACER*

Can Engagement be Compared? Measuring Academic Engagement for Comparison

Ling Tan
Australian Council for Educational Research
19 Prospect Hill Rd,
Camberwell, VIC, Australia 3124
Ling.Tan@acer.edu.au

Xiaoxun Sun
Australian Council for Educational Research
19 Prospect Hill Rd,
Camberwell, VIC, Australia 3124
Xiaoxun.Sun@acer.edu.au

Siek Toon Khoo
Australian Council for Educational Research
19 Prospect Hill Rd,
Camberwell, VIC, Australia 3124
SiekToon.Khoo@acer.edu.au

ABSTRACT
Student engagement is a reflection of active involvement in learning. In digital learning environment, research studies on engagement have been focused on detecting behavioral and psychological engagement indicators from the patterns of activities using feature engineering, but student engagement estimates were rarely compared across sessions or across domains of learning. This paper describes how this could be done by revisiting engagement instrument, diagnosing engagement indicators, estimating engagement parameters, and equating. This study illustrates how engagement reliability can be improved by refining engagement indictors. We demonstrated through DataShop data that student engagement levels can be compared across domains of learning..

Keywords

Behavior, academic engagement, measurement, ITS.

1. INTRODUCTION
In digital learning environment, research study of engagement often focused on detecting behavioral engagement indicators [3,4,5] and psychological engagement indicators [2, 6, 14] using non-intrusive and unobtrusive means. Rather than using surveys to understand engagement, behaviors and affective indicators have been predicted from patterns of activities using feature engineering. The role of machine learning and data mining techniques is to predict behavior or affective status on big data using models developed from training data labeled by human observers. For example, disengagement is inferred by gaming [3, 4] or response time [7]; persistence could be observed by number of revisits to challenging or incomplete tasks [6]; self-regulation could be inferred by the consistency of task completion [1]; and affect status learned from Bayesian Networks [2].

Index of student engagement has been extensively studied to investigate its relationship with learning outcomes. For example, Pardos et.al [14] investigated how well affect states predicted by affect detectors while students worked on exercises throughout a school year in a web-based tutoring platform were correlated with learning outcomes at the end of year. In addition, Rowe, Shores, Mott and Lester [15] found a strong positive relationship between engagement and learning outcomes in narrative-centered learning environments.

This paper is organized as follows. The first section presents the definition of student engagement with a focus on the type of engagement typically found in ITS. The second section presents validity and reliability of academic engagement instrument, diagnostic features of engagement indicators. The last section demonstrates through DataShop data that student engagement levels can be compared across domains of learning.

2. STUDENT ENGAGEMENT CONSTRUT
We argue that there are substantive benefits to study student engagement using methodology found in developing instruments in educational psychology. This approach from instrument point of view offers a number of benefits. Firstly, it sets out to clearly define what kind of student engagement is to be measured at the very beginning. Secondly, it facilitates the comparison of student engagement level across sessions and domains of learning. This means that a student engagement level at the beginning of semester could be compared to the engagement at the middle of semester; and also one's engagement level on Mathematics can be compared to his/her engagement level on Science. Lastly, engagement estimate would be useful for secondary analysis, e.g. correlation between engagement and learning outcomes, or factors influencing engagement which leads to positive learning gains.

2.1 Academic Engagement Construct
It is necessary to develop a valid and reliable measure of student engagement in order to understand the relationship between student engagement and learning outcomes, and to provide tailored strategies to improve learning outcomes of students. Is it possible to define a blue-print of engagement levels in ITS environment like what we would see in conventional self-report survey instrument? The following section will address this issue.

Table 1 provides a preliminary definition of student engagement by levels and corresponding indicators from observed behavioral activities. The definition of student engagement is based on Skinner and Belmont [17], Bomia et al [8], Schlechty [16], Chapman [9], Markwell [13], Willms [18] and Kember, Biggs and Leung [11], and adapted to the indicators in digital learning environment, drawing on additional works by Baker and colleagues [4,5].

Table 1: Mapping of engagement levels to engagement indicators

Level	Behavior Indicators
Level 5: Enthusiasm in learning	Work on additional tasks. Respond to others' questions in online forum. Multiple solutions on tasks.

Level 4: Persistency	Revisiting and spent more time to more difficult tasks. Appropriate use of hints. Completion all tasks. Completion on time.
Level 3: Participation	Work on moderately challenging tasks. Completion of minimum number of tasks.
Level 2: Passive participation	Guessing on majority of tasks. Incompletion on all or majority of tasks. Frequent but inappropriate use of hints.
Level 1: Withdrawal	No response on assignments.

2.2 Data Sets

We used 'Assistments Math 2005-2006' and 'Geometry Area (1996-97)' data sets from PSLC DataShop, available at http://pslcdatashop.org [12]. Both data sets were used for analysing student engagement. The first data set (or Algebra data set) contains action logs of 3136 students using ASSISTments Math tutor from middle schools in a city in central Massachusetts in 2005-2006. Students may use the software for two hours, twice a week. This data set contains 834 unique problems, 2,514 unique steps, total 685,615 transactions of attempting to answer questions and/or requesting helps, and total 6,395 student hours. The data set contains a variety of problem classifications (aka knowledge component).

The second data set (or Geometry data set) is a much smaller data set. This data set was used to compare engagement levels found in the first data set. It has action log data of 59 students using Cognitive Tutor for a Geometry course on a single day, 01/Feb/1996. This data set contains 40 unique problems, 139 unique steps, total 6,778 transactions of attempting to answer questions and/or requesting helps, and total 21 student hours. The data set also contains a variety of Geometry knowledge component classifications in Geometry. Cognitive Tutor system determines which skills a student is having difficulty with, and presents each student with tasks of a skill that he or she has difficulty with. In particular, it estimates the probability of a student knowing each skill based on his/her responses recorded in the system, using Bayesian knowledge-tracing [10].

3. RESULTS

Our first research question is whether it is possible to create an academic engagement instrument guided by engagement construct blueprint outlined in Table 1 from action log data typically recorded in ITS.

3.1 Validity and Reliability

We adapted Baker's behavioral classification [5, 6] and extended it into 11 categories in ITS environments: off-task, gaming, guessing, on-task, on-task using appropriate hints, completion minimum work, completion on time, revisit of moderate-difficult tasks, revisit of hard tasks, extra-task, and extra-time. The extended behavioral classification provides a number of indicators to capture moderate to high levels of academic engagement.

The first 5 behavioral indicators are defined at problem level. *Off-task* is defined as no observations on last-*n* temporal-order tasks, or a student is not working on (or skip) some of assigned tasks. *Gaming* is defined as using excessive hints in a short period of time. *Guessing* is defined as going through difficult tasks quickly without using hints, or going through easy tasks without even spending time reading tasks. *On-task* is defined as working on tasks by producing valid responses after spending a minimum amount of time. *On-task using appropriate hints* is defined as on-task while seeking hints on tasks which are moderately hard relative to student's ability.

The remaining 6 behavioral indicators can be defined at any pre-defined session or mini-session level which contains *n* temporal-order problems. *Completion minimum work* is an indicator to show if a student is able to complete minimum assigned tasks in a session. *Completion on time* is an indicator to show if a student is able to complete minimum assigned tasks on time in a session. *Revisit of moderate-difficult tasks* is set to yes if a student took opportunities to revisit the moderately challenging tasks. *Revisit of hard tasks* is set to yes if a student made an additional efforts to attempt challenging tasks. *Extra-task* indicates if a student made additional efforts to practice on tasks beyond minimum requirement. *Extra-time* indicates if a student spent additional time on assignments.

Behavioral indicators including gaming, guessing, on-task, on-task with appropriate hints, revisit of moderate-difficult tasks, and revisit of hard tasks rely on a critical piece of information, i.e. the likelihood of success on a task. For example, guessing occurs when one finds a particular multiple-choice task hard, and it occurs to students of all ability levels. We can reasonably predict if a student is going to guess if we know the likelihood of success of this student on a particular task.

Prior to estimate student engagement levels of behavioral indicators, observations were arranged in temporal order. For Algebra data set, behavioral indicators were created according to problem-level behavior classifications for *n* problems, which were named as B1 to B*n*. In our experiment, *n* was chosen to be a number close to the average number of problems students attempted in a session (i.e. *n*=12). In addition, six indicators, i.e. completion minimum work, completion on time, revisit of moderate-difficult tasks, revisit of hard tasks, extra-task, and extra-time, were created at session level based on action logs from these *n* problems. ACER ConQuest software [19] was used to estimate KC difficulties and person ability, and the probability of success for each person on each KC was then calculated in SPSS.

What engagement levels are typically found in elements of this instrument? Are the rank orders of instrument indicators working as expected? Figure 1 shows variable map for Algebra data set. The engagement indicators represented by B1 to B12 and the names of six other indicators are displayed on the right hand side of map. The latent engagement levels of individuals represented by "X" are shown on the left hand side. The number of cases represented by each "X" is indicated at the bottom of the variable map. Students at the top of the distribution have higher engagement estimates, while engagement indicators at the top end require higher level of efforts.

The variable map shows that it takes an increasing amount of time or efforts for students to complete more tasks, as indicated by increasing rank order of B1 to B12 in the map. It also shows that it takes more efforts to complete minimum tasks on time than just

to complete minimum tasks. Students who put additional efforts on revisiting challenging tasks, investing more time, or working harder on extra tasks are shown to be more engaged than those who just completing minimum tasks.

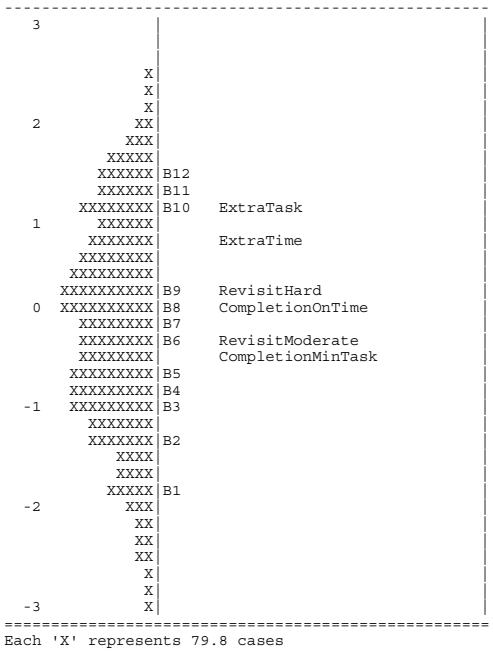

Figure 1: Engagement Variable Map for Algebra Data Set

The reliability coefficients of academic engagement instrument on Algebra data set and on Geometry data set measured by Cronbach's alpha are 0.93 and 0.94 respectively, suggesting correlations among 12 temporal-ordered behavioral indicators and 6 session-level behavioral indicators are high. In conventional survey instruments, reliability coefficients of 0.7 and higher are considered to be reliable. This indicates that reliability of engagement found in these two data sets is as good as those found in conventional survey instruments.

It had been perceived that the rank order of problem-level indicators would be off-task, gaming, guessing, on-task, and on-task using appropriate hints, ordered from the lowest level of engagement to the highest level. We checked this hypothesis by reviewing each indicator. Our detailed analysis on each indicator shows that all problem-level classifications appear to be working as expected, except for on-task using appropriate hints. The rank order of off-task (coded as 0), gaming (coded as 1), guessing (coded as 2), and on-task (coded as 3) can be observed by a clear pattern of increasing average engagement scores. Take the indicator B12 for example (see Table 2). The average engagement scores for off-task cohort, gaming cohort, guessing cohort, and on-task cohort are -0.77, 0.23, 0.57, and 1.32, respectively. However, on-task using appropriate hints did not turn out to have a straight-forward interpretation. In terms of average engagement score, the cohort of on-task using appropriate hints was similar to gaming cohort in observations 1 to 3; similar to guessing cohort in observations 5 to 7; and similar to on-task cohort in observations 10 to 12. For this particular example (i.e. B12), this suggests that it might be better off to combine on-task using appropriate hints with on-task.

Table 2: Engagement Indicator for B12 in Algebra Data Set

Score	Count	% of Total	Pt Bis	Avg	SD
0	9673	68.1	-0.68	-0.77	1.05
1	344	2.4	0.06	0.23	0.53
2	936	6.6	0.18	0.57	0.48
3	2861	20.1	0.61	1.32	0.56
4	392	2.8	0.13	0.88	0.64

3.2 Comparison of Engagement

We have demonstrated validity and reliability of academic engagement instrument through empirical evidence. However, whether the instrument is able to compare engagement levels of a cohort working in Algebra problems with a different cohort working in Geometry problems remains unanswered. In attempting to measure the difference in engagement levels between two different cohorts in different learning contexts of ITS, we will need to create exactly the same behavioral engagement indicators in these two data sets.

Figure 2 shows the scatter plot of behavioral indicator estimates of Algebra data set and indicator estimates of Geometry data set, after adjusting difference in average indicator estimates and ratio of standard deviations. The chart shows that all behavioral indicators had similar rank order in both data sets after taking into account of standard error of estimates. It shows all behavioral indicators were falling into confidence interval lines, except for the indicator, *Revisit of hard tasks* (as circled in red) This indicator appears to be requiring significantly more efforts in Geometry data set than in Algebra data set, with indicator estimates of 0.4 logit in Algebra data set and 1.4 logit in Geometry data set. When the indicator of *Revisit of hard tasks* was excluded, the goodness of fit (R^2) had been significantly improved from 0.78 to 0.99. This indicator was not used in equating due to its large difference in engagement estimates.

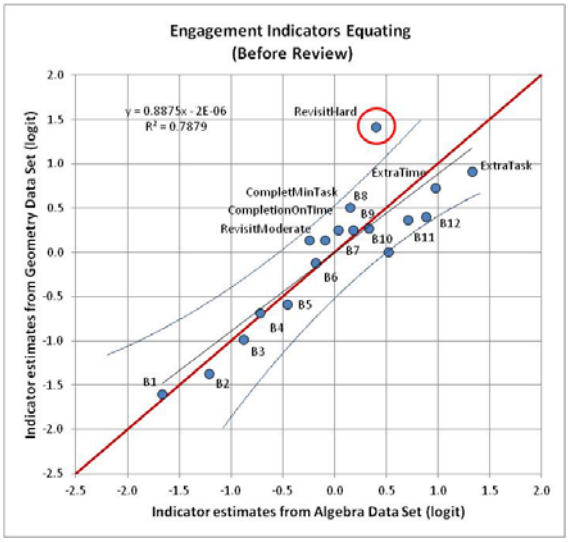

Figure 2: Equating of Engagement Indicators between Algebra Data Set and Geometry Data Set

After applying equating transformation to original engagement scores in Geometry data set, we obtained engagement scores of Geometry data set which can be directly compared to the scores

of Algebra data set. Table 3 shows mean and standard deviations of behavioral engagement scores found in Geometry data and Algebra data. The difference in mean engagement between Geometry and Algebra is 0.225 logit, but this difference is not statistically significant (p-value = 0.089), suggesting academic engagement of a cohort working on Geometry tutor was similar to the engagement of the other cohort working on Algebra tutor. The effect size of the difference in average engagement scores is moderate (Cohen's d = 0.19).

Table 3: Comparison of Average Engagement between Algebra Data Set and Geometry Data Set

Behavioral Engagement in Geometry			Behavioral Engagement in Algebra		
N	**Mean**	**SD**	**N**	**Mean**	**SD**
59	0.123	0.992	14206	-0.101	1.340

4. CONCLUSION

This paper compared student engagement across domains of learning found in two sets of DataShop data. Our preliminary results did not find any significant difference in behavioral engagement between two different cohorts working on two ITS tutors.

5. ACKNOWLEDGMENTS

This research was supported by Australian Research Council ARC-SRI: Science of Learning Research Centre (project number SR120300015). The views expressed herein are those of the authors and are not necessarily those of the Australian Research Council. We acknowledge the use of 'Assistments Math 2005-2006 (3136 Students)' dataset and 'Geometry Area (1996-97)' dataset accessed via DataShop [12]. We thank Ryan Baker, Zach Pardos, and Neil Heffernan for the permission to use their data in our research and sharing their data through the massive open online course of Big Data in Education.

6. REFERENCES

[1] Appleton, J.J., Christenson, S.L., Kim, D. and Reschly, A.L. (2006) Measuring cognitive and psychological engagement: Validation of the Student Engagement Instrument. *Journal of School Psychology*, vol.44, 427-445.

[2] Arroyo, I. and Woolf, B. (2005), Inferring learning and attitudes from a Bayesian network of log file data, *Proceedings of the Twelfth International Conference on Artificial Intelligence in Education*, IOS Press, pp.33–40.

[3] Baker, R.S., Corbett, A.T., Koedinger, K.R., Wagner, A.Z. (2004). Off-Task Behavior in the Cognitive Tutor Classroom: When Students "Game The System". *Proceedings of ACM CHI 2004: Computer-Human Interaction*, 383-390.

[4] Baker, R.S.J.d., Corbett, A.T., Koedinger, K.R., Roll, I. (2006). Generalizing Detection of Gaming the System Across a Tutoring Curriculum. *Proceedings of the 8th International Conference on Intelligent Tutoring Systems*, 402-411.

[5] Baker, R.S.J.d. (2007). Modeling and Understanding Students' Off-Task Behavior in Intelligent Tutoring Systems. *Proceedings of ACM CHI 2007: Computer-Human Interaction*, 1059-1068.

[6] Baker, R.S.J.d., D'Mello, S.K., Rodrigo, M.M.T., and Graesser, A.C. (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.

[7] Beck, J.E.(2005). Engagement tracing: Using response times to model student disengagement. In C-K.Looi et al. (Eds). Artificial intelligence in education: supporting learning through intelligent and socially informed technology, 88-95. Amsterdam: IOS Press.

[8] Bomia, L., Beluzo, L., Demeester, D., Elander, K., Johnson, M., & Sheldon, B. (1997). The impact of teaching strategies on intrinsic motivation, Champaign, IL: ERIC Clearinghouse on Elementary and Early Childhood Education. p. 294.

[9] Chapman, E. (2003). Alternative approaches to assessing student engagement rates. *Practical Assessment, Research & Evaluation*, 8(13). Retrieved 7/2/07.

[10] Corbett, A.T. and Anderson, J.R. (1995), Knowledge Tracing: Modeling the Acquisition of Procedural Knowledge. *User Modeling and User-Adapted Interaction*, 4, 253-278.

[11] Kember, D., Biggs, J. & Leung, D. Y. P. (2004). Examining the Multidimensionality of Approaches to Learning through the Development of a Revised Version of the Learning Process Questionnaire, *British Journal of Educational Psychology*, 74, 261-279.

[12] Koedinger, K.R., Baker, R.S.J.d., Cunningham, K., Skogsholm, A., Leber, B., Stamper, J. (2010) A Data Repository for the EDM community: The PSLC DataShop. In Romero, C., Ventura, S., Pechenizkiy, M., Baker, R.S.J.d.(Eds.) *Handbook of Educational Data Mining*. Boca Raton, FL: CRC Press.

[13] Markwell, D. (2007), A large and liberal education': higher education for the 21st century, Melbourne: Australian Scholarly Publishing & Trinity College, University of Melbourne.

[14] Pardos,Z.A., Baker, R.S.J.d., San Pedro, M.O.C.Z., Gowda,S.M., and Gowda, S.M. (2013), Affective states and state tests: Investigating how affect throughout the school year predicts end of year learning outcomes, *Proceedings of the Third International Conference on Learning Analytics and Knowledge*, pp. 117-124.

[15] Rowe,J.P., Shores,L.R., Mott,B.W., and Lester, J.C. (2011). Integrating Learning, Problem Solving, and Engagement in Narrative-Centered Learning Environments. *International Journal of Artificial Intelligence in Education*, 21, 115–133.

[16] Schlechty, P. (1994). Increasing Student Engagement. *Missouri Leadership Academy*. p. 5.

[17] Skinner, E. A. and Belmont, M. J. (1993). Motivation in the classroom: Reciprocal effects of teacher behavior and student engagement across the school year. *Journal of Educational Psychology*. 85(4), 571-581.

[18] Willms, J.D. (2003). Student Engagement at School: a sense of belonging and participation: Results from PISA 2000. Organisation for Economic Co-operation and Development. p. i.

[19] Wu, M. L., Adams, R. J., Wilson, M. R., Haldane, S.A. (2007). ACER ConQuest Version 2: Generalised item response modelling software [computer program]. Camberwell: Australian Council *for* Educational Research.