Confrustion and Gaming while Learning with Erroneous Examples in a Decimals Game

Michael Mogessie[1], J. Elizabeth Richey[1][0000-0002-0045-6855], Bruce M. McLaren[1], Juan Miguel L. Andres-Bray[2] and Ryan S. Baker[2]

[1] Carnegie Mellon University, Pittsburgh, PA 15213, USA
michaelmogessie@cmu.edu

[2] University of Pennsylvania, Philadelphia, PA 19104, USA

Abstract. Prior studies have explored the potential of erroneous examples in helping students learn more effectively by correcting errors in solutions to decimal problems. One recent study found that while students experience more confusion and frustration (confrustion) when working with erroneous examples, they demonstrate better retention of decimal concepts. In this study, we investigated whether this finding could be replicated in a digital learning game. In the erroneous examples (ErrEx) version of the game, students saw a character play the games and make mistakes, and then they corrected the characters' errors. In the problem solving (PS) version, students played the games by themselves. We found that confrustion was significantly, negatively correlated with performance in both pretest ($r = -.62$, $p < .001$) and posttest ($r = -.68$, $p < .001$) and so was gaming the system (pretest $r = -.58$, $p < .001$, posttest $r = -.66$, $p < .001$). Posthoc (Tukey) tests indicated that students who did not see any erroneous examples (PS-only) experienced significantly lower levels of confrustion ($p < .001$) and gaming ($p < .001$). While we did not find significant differences in post-test performance across conditions, our findings show that students working with erroneous examples experience consistently higher levels of confrustion in both game and non-game contexts.

Keywords: Digital Learning Game, Erroneous Examples, Affect, Affect Detection, Confusion, Frustration, Gaming the System, Learning Outcomes

1 Introduction

Researchers have investigated the value of solving problems using non-traditional approaches to problem solving. Worked examples [1-3] and erroneous examples [4-6] have been of particular interest. Worked examples demonstrate a procedure to arrive at a correct solution and may prompt students to provide explanations to correct steps of a solution while erroneous examples require them to identify and fix errors in incorrect solutions. The reason these approaches improve learning has been attributed to their role in freeing up cognitive resources that can then be used to learn new knowledge [7]. Factors not specific to a particular approach may also interact with learning. Of these, affect and behavior have garnered the most attention [8-11]. In

particular, states of confusion, concentration and boredom have been shown to persist across computer-based learning environments (dialog tutors, problem-solving games, problem-solving intelligent tutors) [12].

In a recent study, we found that students who were assigned erroneous examples implemented in an intelligent tutor [13] experienced higher levels of confrustion [14], a mix of confusion and frustration, than those who were asked to answer typical problem-solving questions. However, we found that confrustion was negatively correlated with both immediate and delayed learning, albeit less so for students who worked with erroneous examples.

This study, which is a replication of our recent findings but in a game versus ITS context, was motivated by two observations. First, in order to determine whether this relationship is robust, it is important to explore whether our recent findings persist in other digital learning environments. This is because levels of affective states such as frustration and behaviors such as gaming the system have been shown to vary across learning environments and user interfaces [15, 12].

Second, research has shown that students who engage in gaming the system also experience frustration [10], though frustration does not always precede gaming [12]. Therefore, it is interesting to explore if this association persists when erroneous examples are implemented in a digital learning game context.

Participants were divided into four groups where two groups worked with either Erroneous Examples (ErrEx) or Problem Solving (PS) questions only and the other two worked with a mix of either ErrEx then PS or PS then ErrEx questions. We expected that students in all four groups would perform better from pretest to posttest. We then tested the following hypotheses:

H1: Confrustion and gaming will be negatively related to performance, even when controlling for prior knowledge.

H2: Students in any of the conditions that include erroneous examples will experience higher levels of confrustion and gaming the system.

H3: Students in any of the conditions that include erroneous examples will perform better than their PS-only counterparts in the posttest.

2 Methods

The data used in this study was collected in the spring of 2015. Participants were recruited from four teachers' classes at two middle schools, and participated over four to five class sessions. Both schools are located in the metropolitan area of a city in the United States. The analysis for this study included the data of 191 students, divided into four conditions within the game context.

Materials consisted of the digital learning game, Decimal Point [16], and three isomorphic versions of a test administered as a pretest and posttest. The Decimal Point game is laid out on an amusement park map, with 24 mini-games in which students play two rounds of each. All tests and the game used the Cognitive Tutor Authoring Tool (CTAT) [17] as a tutoring backend. The game was designed with focus on common misconceptions middle school students have about decimals [18].

We used gameplay data to generate machine learning models to detect confrustion and gaming the system. In this study, we applied text replay coding [19, 20] to student logs to label 1,560 clips (irr κ=.74). To predict confrustion and gaming, the detectors used 23 features of the students' interaction with the decimal tutor, involving the number of attempts, amount of time spent and restart behavior.

After evaluating the performance of several classification algorithms in terms of Area Under the Receiver Operating Characteristic Curve (AUC ROC) and Cohen's Kappa (κ), we built the confrustion detector using the Extreme Gradient Boosting (XGBoost) ensemble tree-based classifier [21] (AUC ROC = .97, κ = .81) and the gaming detector using the J-Rip classifier [22] (AUC ROC = .85, κ = .62)

3 Results

Confrustion was significantly, negatively correlated with performance on the pretest (r = -.62, p < .001) and posttest (r = -.68, p < .001). A multiple regression model tested using confrustion to predict posttest performance while controlling for pretest was also significant, $F(2, 188)$ = 181.14, p < .001. Within the model, both pretest, (β = .57, p < .001) and confrustion (β = -.32, p < .001) were significant; confrustion was a significant, negative predictor of posttest performance even after controlling for pretest.

Gaming was significantly, negatively correlated with performance on the pretest (r = -.58, p < .001) and posttest (r = -.66, p < .001). A multiple regression model tested using gaming to predict posttest performance while controlling for pretest was also significant, $F(2, 188)$ = 181.14, p < .001. Within the model, both pretest, (β = .59, p < .001) and gaming (β = -.31, p < .001) were significant, indicating that gaming was also a significant, negative predictor of posttest performance even after controlling for pretest.

Mean levels of confrustion and gaming for each condition are reported in Table 1. A one-way analysis of variance (ANOVA) comparing gaming and confrustion levels across conditions indicated a significant effect of condition on confrustion, $F(3, 187)$ = 14.01, p < .001, and gaming, $F(3, 187)$ = 10.07, p < .001. Posthoc (Tukey) tests indicated that students in the PS-only condition experienced significantly lower levels of confrustion (ps < .001), while there were no differences among the other conditions (ps > .97). Similarly, posthoc (Tukey) tests indicated that students in the PS-only condition experienced significantly lower levels of gaming(ps < .001), while there were no differences among the other conditions (ps > .91).

Table 1. Gaming, confrustion, and test performance by condition.

Measure	PS	ErrEx	ErrEx/PS	PS/ErrEx
Pretest (*SD)*	23.37 (8.20)	23.39 (9.05)	20.92 (8.00)	20.48 (8.41)
Posttest *M* (*SD*)	28.72 (6.18)	27.63 (7.61)	25.75 (7.15)	26.40 (7.43)
Gaming *M* (*SD*)	.16 (.11)	.27 (.17)	.28 (.13)	.29 (.13)
Confrustion *M* (*SD*)	.24 (.16)	.46 (.26)	.45 (.20)	.47 (.18)

Finally, a repeated-measure analysis of variance (ANOVA) indicated that students across all conditions improved significantly from pretest to posttest, $F(3, 187) = 167.04$, $p < .001$. See Table 1 for means and standard deviations across conditions. A series of ANOVAs indicated no significant differences across conditions on pretest, $F(3, 187) = 1.63$, $p = .18$, or posttest, $F(3. 187) = 1.65$, $p = .18$.

4 Discussion

In this study, we implemented erroneous examples in a digital learning game context and found that students who played the erroneous examples versions of the game experienced higher levels of confrustion. There was also a significant correlation between gaming the system and confrustion. Future research might further explore the relationship between frustration and gaming, as previous research using affect detectors has found that frustration did not tend to precede gaming the system [12].

A previous study using a web-based intelligent tutor showed that students working with erroneous examples performed better than their problem-solving counterparts [23]. This study, however, did not replicate that finding.

While it is not possible to make a direct comparison between confrustion levels in the game and intelligent tutor versions of the ErrEx condition, it is worth noting that students who played the game experienced higher levels of confrustion (M = 0.46, SD = 0.26) than those who used the intelligent tutor (M = 0.34, SD = 0.16) [13]. Since confrustion has been shown to be significantly, negatively correlated with learning, these higher levels of confrustion may explain why we did not see better learning effects of erroneous examples in the game context.

Alternatively, integrating the game interface with a feature where students watch a game character play the game for them may have negatively impacted both the game experience and the intended benefit of erroneous examples.

In an upcoming study, we will explore mechanisms intended to reduce the negative impact of confrustion and gaming on learning with erroneous examples in a digital learning game.

References

1. Sweller, J., Cooper, G.A.: The use of worked examples as a substitute for problem solving in learning algebra. Cognition and Instruction, 2, 59-89 (1985)
2. McLaren, B.M., Lim, S., Koedinger, K.R.: When and how often should worked examples be given to students? New results and a summary of the current state of research. In: Proceedings of the 30th Annual Conference of the Cognitive Science Society, (pp. 2176-2181). Austin, TX: Cog. Sci. Society (2008)
3. Renkl, A., Atkinson. R.K.: Learning from worked-out examples and problem solving. In. J.L. Plass, R. Moreno, & R. Brünken (Eds.), Cognitive Load Theory, Cambridge: Cambridge University Press (2010)
4. Isotani, S., Adams, D., Mayer, R.E., Durkin, K., Rittle-Johnson, B., McLaren, B.M.: Can erroneous examples help middle-school students learn decimals? In:

Kloos, C.D., Gillet, D., Crespo García, R.M., Wild, F., Wolpers, M. (eds.) EC-TEL 2011. LNCS, vol. 6964, pp. 181–195. Springer, Heidelberg (2011).

5. Adams, D., et al.: Using erroneous examples to improve mathematics learning with a web-based tutoring system. Comput. Hum. Behav. 36C, 401–411 (2014)

6. McLaren, B.M., Adams, D.M., Mayer, R.E.: Delayed learning effects with erroneous examples: a study of learning decimals with a web-based tutor. Int. J. Artif. Intell. Educ. 25 (4), 520–542 (2015)

7. McLaren, B.M., Adams, D., Durkin, K., Goguadze, G. Mayer, R.E., Rittle-Johnson, B., Sosnovsky, S., Isotani, S. & Van Velsen, M. (2012). To err is human, to explain and correct is divine: A study of interactive erroneous examples with middle school math students. In: A. Ravenscroft, S. Lindstaedt, C. Delgado Kloos, & D. Hernándex-Leo (Eds.), Proceedings of ECTEL 2012: Seventh European Conference on Technology Enhanced Learning, LNCS 7563 (pp. 222-235). Springer, Berlin.

8. Prensky, M. (2003). Digital game-based learning. Computers in Entertainment (CIE), 1(1), 21-21.

9. 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". In Proceedings of the SIGCHI conference on Human factors in computing systems (pp. 383-390).

10. Walonoski, J. A., & Heffernan, N. T. (2006). Detection and analysis of off-task gaming behavior in intelligent tutoring systems. In International Conference on Intelligent Tutoring Systems (pp. 382-391). Springer, Berlin, Heidelberg.

11. Gee, J. P. (2012). Situated language and learning: A critique of traditional schooling. routledge.

12. Baker, R. S., D'Mello, S. K., Rodrigo, M. M. T., & 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.

13. Richey, J. E., Andres-Bray, J. M. L., Mogessie, M., Scruggs, R., Andres, J. M., Star, J. R., ... & McLaren, B. M. (2019). More confusion and frustration, better learning: The impact of erroneous examples. Computers & Education, 139, 173-190.

14. Liu, Z., Pataranutaporn, V., Ocumpaugh, J., & Baker, R. (2013). Sequences of frustration and confusion, and learning. In Educational data mining 2013.

15. Baker, R. S., De Carvalho, A. M. J. A., Raspat, J., Aleven, V., Corbett, A. T., & Koedinger, K. R. (2009). Educational software features that encourage and discourage "gaming the system". In Proceedings of the 14th international conference on artificial intelligence in education (pp. 475-482).

16. McLaren, B. M., Adams, D. M., Mayer, R. E., & Forlizzi, J. (2017). A computer-based game that promotes mathematics learning more than a conventional approach. International Journal of Game-Based Learning (IJGBL), 7(1), 36-56. doi:10.4018/IJGBL.2017010103

17. Aleven, V., McLaren, B.M., Sewall, J.: Scaling up programming by demonstration for intelligent tutoring systems development: an open-access

website for middle school mathematics learning. IEEE Trans. Learn. Technol. 2(2), 64–78 (2009).

18. Isotani, S., McLaren, B.M., Altman, M.: Towards intelligent tutoring with erroneous examples: A taxonomy of decimal misconceptions. In: International Conference on Intelligent Tutoring Systems. pp. 346–348. Springer (2010).

19. Lee, D. M. C., Rodrigo, M. M. T., d Baker, R. S., Sugay, J. O., & Coronel, A. (2011). Exploring the relationship between novice programmer confusion and achievement. In International conference on affective computing and intelligent interaction (pp. 175-184). Springer, Berlin, Heidelberg.

20. Baker, R. S., Corbett, A. T., & Wagner, A. Z. (2006). Human classification of low-fidelity replays of student actions. In Proceedings of the educational data mining workshop at the 8th international conference on intelligent tutoring systems (Vol. 2002, pp. 29-36).

21. Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining (pp. 785-794). ACM.

22. Cohen, W.W. 1995. Fast effective rule induction. Twelfth International Conference on Machine Learning. (1995), 115–123.

23. McLaren, B. M., Adams, D. M., & Mayer, R. E. (2015). Delayed learning effects with erroneous examples: a study of learning decimals with a web-based tutor. International Journal of Artificial Intelligence in Education, 25(4), 520-542.