Title
Productive Failure and Student Emotions

Journal
Proceedings of the Annual Meeting of the Cognitive Science Society, 43(43)

ISSN
1069-7977

Authors
Savelson, Zachary M.
Muldner, Kasia

Publication Date
2021

Peer reviewed

Productive Failure and Student Emotions

Zachary Savelson (zachsavelson@cmail.carleton.ca)
Department of Cognitive Science, 1125 Colonel By Drive
Ottawa, ON, K1S5B6, Canada

Kasia Muldner (kasiamuldner@cunet.carleton.ca)
Department of Cognitive Science, 1125 Colonel By Drive
Ottawa, ON, K1S5B6, Canada

Abstract

Productive failure (PF) is a learning paradigm that reverses the standard order of instruction by asking students to solve problems *prior* to instruction. This paradigm has been shown to be effective for fostering student learning. To date, however, the role of student emotion in productive failure has not been investigated. In other paradigms, there is some evidence that failure elicits negative emotions and that these emotions can interfere with learning. This leads to a conundrum given productive failure's positive effect on learning. To shed light on this, we report on results from a study ($N = 48$) in the productive failure paradigm. For the analysis, we used a mixed-methods approach to investigate the distribution of emotions in productive failure, how these changed across different instructional activities, and the relation between emotions and posttest performance.

Keywords: Productive failure; emotions

Introduction

Theories of impasse-driven learning show that failure is an important part of learning (Darabi et al., 2018; Tawfik, Rong & Choi, 2015; VanLehn 1988). Failure can have both negative and positive effects. On the one hand, it can result in students feeling discouraged and frustrated (Falout et al., 2009; Nummenmaa & Niemi, 2004), but on the other hand, it can promote reflection and serve as a catalyst for novel ideas needed to overcome impasses (D'Mello et al., 2014; Kapur & Bielaczyc, 2012; Loibl & Rummel, 2014). The present study focuses on a particular failure paradigm called *productive failure* (Kapur, 2008; Kapur & Bielaczyc, 2012). Our goal was to investigate what emotions productive failure elicits in students and whether those emotions relate to their learning.

Productive Failure

The productive failure (PF) paradigm includes two key phases, with the "failure" part occurring during the first phase. Here, students are given a problem to work on that they do not have adequate prior knowledge to solve (referred to as the *PF phase* below). Not surprisingly, students rarely find correct solutions, and in the process of attempting to solve the problem they encounter impasses and setbacks. However, they also have the opportunity to be creative, try unconventional solutions, and explore the problem space. These activities are hypothesized to help students become aware of their knowledge gaps, activate prior knowledge, and recognize deep conceptual features of the problem (Kapur, 2008). The second productive failure phase (the *Lesson phase*) involves giving students a lesson on how to solve the problem, including the salient concepts related to the solution. This ordering (problem solving first, lesson second) may seem counterintuitive as it is contrary to traditional instruction. However, there are benefits to productive failure: students learn more from this paradigm than traditional instruction (Kapur, 2008; 2012; Kapur & Kinzer, 2009; Loibl & Rummel, 2014). Productive failure is especially beneficial for fostering conceptual knowledge (the 'why / what' understanding of key domain concepts). This is because productive failure is designed to expose students to the salient problem features. In contrast, both productive failure and traditional instruction are similarly effective at fostering procedural knowledge (the 'how to' knowledge - in problem-solving contexts, it can be defined as the ability to execute a series of actions to solve a problem (Rittle-Johnson, Siegler & Alibali, 2001)).

To illustrate with a few examples, the classic productive failure study (Kapur 2012) evaluated the paradigm in a classroom with ninth-grade students. There were two conditions: productive failure and direct instruction. The productive failure group spent the first two class periods working collaboratively in triads to solve a math problem centered around the concept of variance. Subsequently, students were given a lesson by their teacher, who explained the concept of variance and also provided contrasting cases by comparing and contrasting various solution methods (including canonical and incorrect solutions). For the direct instruction condition, the order of instruction was flipped: the first period consisted of the teacher explaining the concept of variance and the next two periods involved problem solving. Both conditions were subsequently given a posttest. The productive failure group significantly outperformed the traditional group on the conceptual and transfer items; there was no difference on the procedural items.

Loibl & Rummel's (2014) study replicated and extended the earlier productive failure studies, while also addressing a limitation in those studies. That limitation was due to a difference between the productive failure and traditional conditions in the lesson, namely that only the productive failure group's lesson included contrasting cases. Contrasting cases involve incorrect solutions used to teach students about the critical problem features. Thus, the fact that only the productive failure group had instruction on contrasting cases could bias the results against the traditional group. To address this limitation, Loibl and Rummel used a 2x2 design, manipulating the lesson design (contrasting cases *present* or *not*) and the timing of the lesson (*before* or *after* problem solving). The participants were 10^{th} graders, and the topic was mathematical variance. Students in the 'failure paradigm' (problem solving first - lesson after) acquired more conceptual knowledge than the traditional instruction group (lesson first-problem solving after). Additionally, the main effect of lesson design was significant, favoring the lesson that included contrasting cases. These results replicate earlier findings that productive failure is an effective paradigm and demonstrates that including contrasting cases increases learning.

To summarize, productive failure fosters greater conceptual learning than traditional instruction, without sacrificing procedural learning. We now review research on emotions and learning.

Emotion and Learning

Research outside of the productive failure paradigm has shown that students' emotions are an integral part of learning and impact outcomes (D'Mello et al., 2014; Kim & Pekrun, 2014; MacIntyre & Vincze, 2017). Certain emotions improve learning. For example, when students enjoy their work and feel happy, they are more creative and flexible with their learning styles (Kim & Pekrun, 2014). MacIntyre & Vincze (2017) reported that enjoyment positively impacted students' second language learning. Confusion can be beneficial when it is induced through contradictory information, leading to a confrontation of the corresponding impasse, and so learning (D'Mello et al., 2014). Uncertainty also positively correlates with learning in some studies (Lamnina & Chase 2019; Ozcelik, Cagiltay, & Ozcelik, 2013).

Conversely, negative emotions like boredom and anxiety have consistently been associated with reduced learning (Craig et al., 2004; Pekrun, Elliot & Maier, 2009). The results related to frustration are mixed: while some models predict a negative relation between frustration and learning (Kort, Reilly & Picard, 2001) others have found the opposite, namely that learning and frustration are positively related but with the caveat that this depends on frustration level (Liu et al., 2013).

In other (non- productive failure) contexts, failure has been correlated with increased anxiety, and, in some cases, hopelessness (Pekrun, Elliot & Maier, 2009). Thus, investigating the emotions elicited during productive failure would shed light on how students feel in this paradigm and provide additional insight into the relationship between emotion and learning in this paradigm.

To our knowledge, to date there is only one study measuring emotion in a productive failure context (Lamnina & Chase, 2019). In this study, middle-school children self-reported on positive and negative emotions after the instructional activity. The results showed that uncertainty and positive affect both increased students' performance on problems. Our work adds to this research by (1) reporting on student emotion during productive failure through a qualitative analysis of verbal protocols, (2) collecting self-reports of emotion at multiple points during the study in order to obtain trends over time. We also used a novel domain, namely computer programming, and an older population, namely university students.

The Present Study

The high-level goal of the present study was to investigate student emotions during the PF phase. We took an exploratory approach by using a single-condition design and measuring emotion through (1) protocol analysis based on transcripts of students working on a productive failure task prior to instruction and (2) self-reports.

Following the productive failure design and guidelines that Kapur & Bielaczyc (2012) proposed, participants in our study worked collaboratively in pairs on a problem. This problem was in the programming domain. This domain facilitates the design of problems according to the productive failure criteria, namely: the problem should allow for exploration of the problem space and be familiar to students on a general level (so that they can use their intuitions to try and solve it). Programming is challenging for many students (Costa & Miranda, 2017) and so novel pedagogical interventions are needed. To date, productive failure studies have focused on math domains and younger populations, and so less is known about outcomes from this paradigm for university students (our population) and programming (our domain). We had the following three research questions:

1. What is the distribution of emotions expressed during productive failure?
2. Do emotions change based on instructional activity?
3. What is the relationship between emotions and posttest performance?

To address these questions, we used both qualitative and quantitative methods. Qualitative methods provided insight into the process of problem solving (PF phase) and in particular the emotions that participants experienced. Quantitative methods provided a complimentary view of the results – here, we used both descriptive and inferential statistics. For the latter, we report both Bayesian and frequentists statistics. While frequentists statistics are the norm, Bayesian statistics provide a measure of evidence present in the data for each model (alternative vs. null) and allow researchers to make claims about the null hypothesis (Jarosz & Wiley, 2014). We report the Bayes factor, where BF_{01}, is "*a ratio that contrasts the likelihood of the data fitting under the null hypothesis with the likelihood of fitting under the alternative hypothesis*" (Jarosz & Wiley, 2014). The inverse, BF_{10}, states the ratio in terms of the alternative hypothesis. We report the more likely model (null BF_{01} or alternative BF_{10}) and interpret the strength of the evidence for that model using Table 4 in Jarosz and Wiley (2014).

Materials

We used the web-based SNAP! programming environment (snap.berkeley.edu) – see Figure 1 for an example. SNAP! does not require syntax knowledge because it provides blocks of code that participants snap together. We customized the SNAP! interface by creating a template so that the only blocks

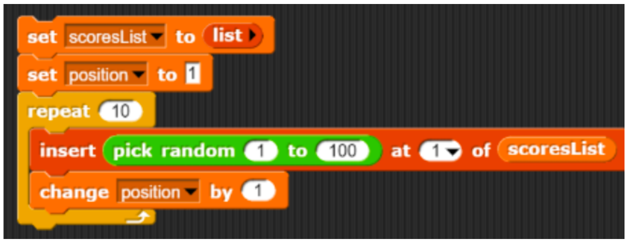

Figure 1: Example of SNAP! program from the tutorial

participants were able to see/use were those relevant to the task. To help participants navigate SNAP!, we created a SNAP! cheat sheet, including labels for the interface components to help the participants guide their collaboration.

Productive Failure Task and Instructional Videos The task used during the PF phase corresponded to sorting a list of numbers – this task meets the criteria for a productive failure task (Kapur & Bielaczyc, 2012) – here, we used a version of the selection sort algorithm. To prepare students for the sorting task, we developed a *background tutorial video* to provide background on programming constructs (variables, loops, conditional statements) and how to implement them in SNAP! Note that this video was not about sorting (and so did not describe it) but rather the foundations needed to approach the sorting task.

A standard productive failure paradigm involves a lesson that follows the problem-solving task. Accordingly, we also created a *sorting lesson video* that had two parts: (1) conceptual foundations, which introduced the sorting algorithm and provided contrasting cases on the key concepts underlying the algorithm, and (2) a step-by-step demonstration that showed participants how to implement the sorting algorithm in SNAP!

Pretest and Posttest The pretest consisted of two procedural questions: (1) a code tracing question and (2) a code-generation question. These questions were designed to test participants' procedural knowledge; the first tested ability to read code and the second tested ability to write code. The posttest, which measured procedural and conceptual knowledge, consisted of three sections: (1) the two pretest questions (*procedural* section), one question asking participants to replicate the sorting algorithm (*recall* section) (to gauge retention from the experiment), and seven questions on the concepts behind the sorting algorithm (*conceptual* section) (to measure participants' conceptual understanding related to sorting). The posttest included more questions than the pretest due to participants' lack of prior programming knowledge at pretest. While having an identical pre and posttest would give us more sensitivity, asking participants about sorting on the pretest would not have been productive, given that sorting is a more advanced topic.

Questionnaires A basic questionnaire measured demographic information. To measure self-reported emotion during the study, we used an emotion self-report questionnaire based on the one in (Muldner et al., 2015) that we adapted for the present study. This questionnaire consisted of five questions, one per target emotion, asking participants how they felt (e.g., *How anxious are you?)*. Responses were provided using a five-point Likert scale (1 = *not at all* … 5 = *extremely*). The emotions in the questionnaire included anxiety, boredom, enjoyment, frustration, and confusion. We chose these emotions because prior research has shown them to be present in learning situations (Craig et al., 2004; D'Mello et al., 2014; Liu et al., 2013; MacIntyre & Vincze, 2017; Pekrun, Elliot & Maier, 2009). We did not have participants report on a wider range of emotions to avoid fatigue.

Participants

The participants were 48 university students (22 males, 21 females, 5 chose not to answer) between the ages of 17 to 29 (M = 21, SD = 3.26). To be eligible, participants had to have (1) *some* self-reported knowledge of computer programming (from tinkering, from a high school course, or from one university course) but (2) not too much knowledge (no more than one university programming class) and (3) come to the study with a friend (as the study involved working with one other individual during the problem-solving task). Thus, in total 24 pairs of students participated.

Design and Procedure

As noted above, we used a single-group design, with all participants assigned to the productive failure group, i.e., they first solved a problem collaboratively (PF phase) and then received a lesson. Each study session was conducted with one pair of participants over a two-hour period using Zoom; all activities were done online. After providing consent, participants individually completed several questionnaires (they are not part of the results here and so are not described) and then the pretest. Following the pretest, the experimenter shared their screen with both participants and showed the background tutorial video. To control for time and keep the experiment a reasonable length, participants could not manipulate the video (e.g., pause).

Following the tutorial, the PF phase began, during which participants were asked to work collaboratively to sort a list of numbers using SNAP! (20 minutes). To help with SNAP! logistics, participants were given a template with a list of randomly generated numbers and the SNAP! cheat sheet. One participant shared their SNAP! screen and was in charge of entering the program steps, while the other participant contributed by specifying actions and ideas. This type of scenario was necessitated by the online nature of the study[1] but is not uncommon in in-person contexts (e.g., students sharing a computer during a lab, with one student in charge

[1] It was not appropriate to ask one of the participants to give remote access to their computer to their partner, because remote access is risky if not done correctly.

of entering the solutions). This phase was screen and audio recorded. Next, participants individually answered the emotion self-report questionnaire. They then watched on the experimenter's shared screen the sorting lesson video that detailed the solution to the problem. Thus, we implemented the productive failure paradigm by having students first work on a problem (phase1) and then receive a lesson on it (phase 2). Following the lesson, participants individually answered the emotion self-report questionnaire, completed the posttest, and answered the self-report questionnaire one more time.

Results

The results are organized according to the research questions presented above.

What is the Distribution of Emotions Expressed During Productive Failure?

To identify the distribution of emotions expressed during the PF phase, we transcribed the audio recordings corresponding to this phase and analyzed them according to guidelines in (Chi, 1997). Our original goal was to identify the same set of emotions as measured in the emotion self-report questionnaire. However, we did not find sufficient evidence of enjoyment, boredom, or anxiety in the verbal protocols. Participants rarely expressed explicitly being in those states nor did they provide sufficient evidence to objectively code for these emotions. Thus, the final coding scheme includes emotions that emerged from the transcripts – these emotions include *frustration*, *confusion*, *uncertainty,* and *positive affect*. We also coded for impasses. All of these emotions/states have been identified in prior studies as present during learning (D'Mello et al., 2014; MacIntyre & Vincze, 2017; Lamnina & Chase, 2019; Liu et al., 2013; VanLehn, 1988). We included both uncertainty and confusion rather than collapsing them into a single code to obtain a finer-grained view on when participants had a mental schema of the problem that did not account for a novel piece of information/development and was trying to reconcile the two (confusion) vs. when participants had an idea but expressed uncertainty about it.

To accurately capture the presence of the target states, we developed an initial coding scheme. Two researchers independently coded a series of 4 transcripts; after each transcript was coded, disagreements were discussed, and the scheme was updated. After 4 transcripts, saturation was reached (the scheme stabilized). To obtain inter-rater agreement, 5 more transcripts were independently coded (no discussion occurred during this period). On this set, Cohen's Kappa was .681, $p < .01$, indicating substantial agreement between the coders (McHugh, 2012). The finalized coding scheme, shown in Table 1, was used by the primary researcher to code the remaining transcripts.

The mean number of times the target emotions were expressed per participant is shown in Figure 2. Excluding "other" utterances, 62% corresponded to uncertainty (496), 15% to confusion (119), 16% to impasses (130), 3.5% to frustration (30), and 3.5% to positive affect (28).

Table 1: Scheme used to code participants' transcripts

Uncertainty	Participants expressed they were not sure about an idea's validity, as indicated by hedges, hypotheticals, suppositions, or probability statements (Jordan et al., 2014).
Confusion	Participants expressed confusion, posed a 'why' question or expressed that an idea or result was not what they expected.
Frustration	Participants expressed frustration, either as a response to confusion, uncertainty, or impasse, or a general expression of irritation.
Positive Affect	Participants expressed any positive emotion, including success-related utterances (e.g., "we did it!").
Impasse	Participants expressed they did not know how to proceed. Explicit expressions of 'being stuck' also counted as impasses.
Other	All other utterances (including fragments, simple agreement, task coordination, shallow questions like "what did you say?")

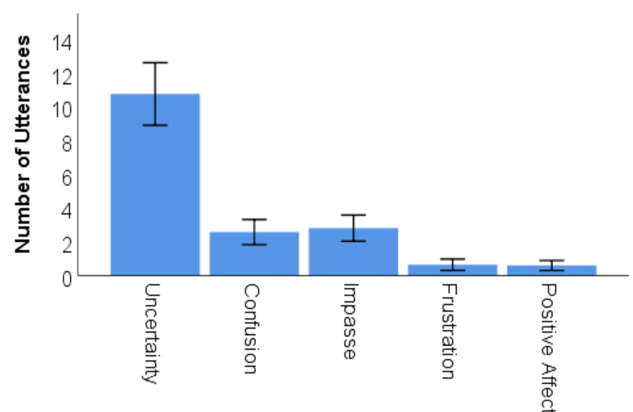

Figure 2: Mean number of times emotions and impasses were expressed per participant (error bars = 95% CI)

Do Emotions Change Based on Instructional Activity?

To examine participants' self-reported emotions after each key activity (PF phase – sorting task, lesson, and posttest), we used data from the self-report questionnaire that was collected at these three timepoints. We did not include uncertainty in the self-report questionnaire because based on our coding scheme's definition (see Table 1), uncertainty had to be related to a particular idea whereas these 5 emotions are not bound to an idea. Figure 3 shows the trends in how emotions varied over time for each of the five target emotions.

As reported by a one-way repeated measures ANOVA for each emotion, there was a significant main effect of time point on emotion (anxiety: $F(2,94) = 11.01$ $p < .01$, $\eta_p^2 = .19$; boredom: $F(2,94) = 29.2$, $p < .01$, $\eta_p^2 = .38$; enjoyment: $F(1.7, 80.6) = 5.8$, $p <.01$, $\eta_p^2 = .11$; frustration: $F(2,94) = 15.7$, $p < .01$, $\eta_p^2 = .25$; confusion: $F(1.8,82.4) = 14.9$, $p < .01$, $\eta_p^2 = .24$). Bayesian statistics provided decisive evidence for the alternative model ($BF_{10} > 150$ for each analysis). These results demonstrate that participants' levels of all emotions were affected by the activity that directly preceded the emotion self-report. We used trend analysis to further investigate these patterns over time. All emotions followed a quadratic trend (boredom, anxiety, frustration, confusion: $p < .01$; enjoyment: $p = .053$). For boredom, the trend indicated boredom was lowest directly after PF phase (sorting task), highest after the lesson,and then decreased slightly after the posttest. This suggests that students self-reported more interest after the sorting task compared to the lesson. The other four emotions (anxiety, confusion, frustration, and enjoyment) also followed a quadratic trend, but the pattern was different: each emotion started off high after PF phase but dropped after the lesson, increasing again after the posttest. These results highlight that the PF phase elicits higher emotional activation than the more passive lesson viewing – this is not surprising as students are more actively engaged in the content during productive failure than when viewing the lesson. After the posttest, the level of boredom decreased while the levels of the other four emotions increased. Not surprisingly, these changes highlight that the posttest elicited a greater emotional response than the lesson (although descriptively, not as high an emotion response as was self-reported directly after the PF phase).

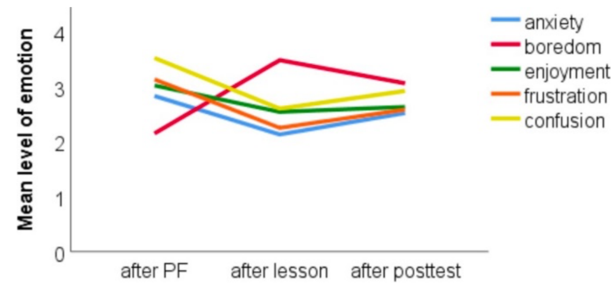

Figure 3: Change in self-reported emotion over the three time points for the five target emotions

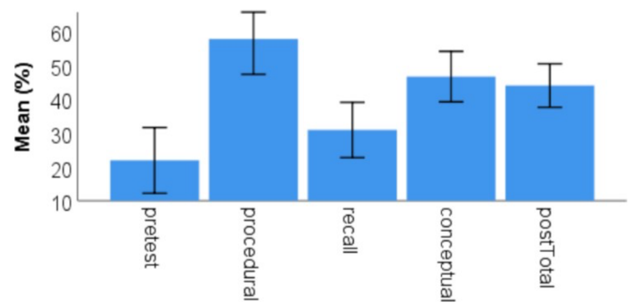

Figure 4: Pretest %, procedural %, recall %, conceptual % and postTotal % (error bars = 95% CI)

What is the Relationship Between Emotions and Posttest Performance?

We begin with the results on programming knowledge, assessed by the pretest and posttest (Figure 4). As noted above, the pretest assessed only procedural knowledge as participants were not expected to have any algorithmic or conceptual knowledge of sorting. The pretest confirms that participants had low prior programming knowledge, with the average pretest score at 22%. The posttest % was also low (44%). Figure 4 also shows the percentage obtained for the three types of questions, procedural, algorithm recall, and conceptual. Of the three types of questions, participants performed the highest on the procedural questions (58%, a 36% increase from the pretest).

We investigated the relationship between posttest performance and emotions using exploratory correlational analysis; the emotion data included (1) the emotions expressed during the PF phase (see Figure 2), and (2) the self-reported emotions directly after the PF phase (see after lesson, Figure 3), namely we correlated each of the target emotions with the total posttest %. We acknowledge this analysis involves a relatively large number of comparisons but given the exploratory nature of the work we felt proceeding was warranted.

We begin with the results for the emotions extracted from the transcripts during the PF phase. The frequentist statistics did not identify any significant correlations between posttest% and target emotion (frustration, confusion, uncertainty, and positive affect), $p > .3$ for all analyses. We also found no relation between any of the emotions and the individual posttest sections (procedural, recall, and conceptual). We turned to Bayesian statistics to see if there was evidence for the null model. This turned out to be the case with substantial evidence for the null model for frustration ($BF_{01} = 4.3$), uncertainty ($BF_{01} = 5.3$), confusion ($BF_{01} = 3.4$), and positive affect ($BF_{01} = 4.9$).

The results for the self-reported emotions directly after the PF phase indicated a significant positive relationship between self-reported enjoyment and posttest % ($r = .32$, $p = .03$). While Bayesian analysis confirmed the alternative model was more likely, the evidence was anecdotal ($BF_{10} = 1.8$). Confusion was negatively associated with posttest % ($r = .28$, $p = .054$) but the Bayesian analysis indicated both models were more or less equally likely ($BF_{10} = .1$; $BF_{01} = .9$). Thus, we cannot draw conclusions about the relation between post-

PF phase confusion and posttest %. For the remaining emotions, no correlations were significant (all $p > .17$) and the Bayesian statistics provided substantial evidence for the null model for two emotions (anxiety $BF_{01} = 4.9$; boredom $BF_{01} = 4.4$). For frustration, the null model was more likely, but the evidence was anecdotal ($BF_{01} = 2.0$).

Discussion

We extended prior work on productive failure by exploring the distribution and change in emotions with university students through both verbal protocol analysis and self-report data, in a novel domain (computer programming). We now discuss our primary findings.

Emotions in Productive Failure

To gain insight into emotions elicited by productive failure, we analyzed the transcripts from the PF phase to identify expressions of frustration, confusion, uncertainty, and positive affect, as well as impasses. We acknowledge, however, that the PF phase may have not been the only cause of the measured emotions. In particular, the pretest could have caused students to feel emotions that they carried over into the PF phase, something we plan to take into account in future work.

While we did find that all five target states were expressed during the PF phase, with the exception of uncertainty the levels of expression were low. Baker et al. (2010) also found low levels of emotional expression during standard problem-solving activities – in this study, emotions were measured by observers and the most common states were engaged concentration (60% of the time) and confusion (13% of the time). We anticipated that productive failure might elicit more expression of confusion, uncertainty, and frustration because students are asked to work on problems they are not expected to succeed in solving. One potential explanation for why this did not happen could be that the collaboration between the participants during the PF phase buffered the negative affect that usually accompanies failure. Pietarinen et al. (2018) reported a positive relationship between collaboration and positive affect. While this finding hints at the possibility that collaboration in productive failure may reduce negative emotion and related states, in our study expression of positive affect was also low. Thus, in general participants were not that verbally expressive about their emotions in the PF phase. Here, additional channels beyond verbal data may help to more accurately measure emotion, such as data from physiological sensors like skin conductance, something we did in prior work but outside the productive failure paradigm (Muldner et al. 2010). Another source of data could correspond to human judges, an approach used in other work (Baker et al. 2010).

In addition to identifying the emotions expressed during the PF phase, we used self-report questionnaires to collect information on emotion levels right after (1) the PF phase, (2) the lesson, and (3) the posttest. Boredom was lowest right after the PF phase compared to after the other two activities, highlighting productive failure's potential to engage students. Confusion, anxiety, and frustration were also higher after the PF phase, with confusion rated slightly higher than the other two emotions. As noted, self-reports were used to supplement the verbal protocol data. One limitation with doing so is that there are individual differences in ability to accurately self-report (Barrett, 2004; Barrett et al., 2004). Thus, future work could supplement the self-reports with additional channels of data (e.g., physiological sensors, human judges).

Posttest Performance and Emotions

Some prior work reported relationships between emotions and learning (Pekrun, Elliot & Maier, 2009). As another example, D'Mello et al. (2014) found that the more confused students were, the better their learning outcomes were. In contrast, we found that emotions were not predictive of posttest performance. We tested this relationship both for emotions expressed during the PF phase and directly after. While frequentist statistics do not allow us to make claims about null results, we also used Bayesian statistics, which do allow for such claims.

One explanation for the lack of association between emotion and posttest scores is that students did not experience emotions that frequently during productive failure in our study. With the exception of uncertainty, expressions of emotion were rare during PF phase and moderate right after the PF phase. It may be that students would need to experience emotion more frequently (and/or strongly) for the emotion to have an effect on learning. Another potential explanation for the lack of a correlation between emotion and learning is that we gave participants a short (20 minute) window to work on the productive failure task. While this was necessary due to experimental nature of the study (and thus the need to control time on task), giving students a longer period of time with the option to stop early if they choose may make the effect of emotion more salient.

Other Future Work

To date studies have focused on productive failure outcomes. Through our analysis of the transcripts captured during the PF phase, our study adds to the productive failure paradigm by providing insight into the productive failure process. In the present study, this analysis is limited to emotional states but as the next step, we are analyzing students' collaborations to identify constructive and interactive patterns. Our plan is to analyze if more constructive collaborations are associated with less negative emotion and higher posttest performance.

Acknowledgements

This work was supported with an NSERC Discovery Grant #1507 and a Masters' SSHRC grant.

References

Baker, R., D'Mello, S. K., Rodrigo, 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.

Barrett L.F. (2004). Feelings or words? Understanding the content in self-report ratings of experienced emotion. *Journal of Personality and Social Psychology*, 87(2), 266-281.

Barrett, L.F., Quigley, K.S., Bliss-Moreau, E., & Aronson, K.R. (2004). Interoceptive sensitivity and self-reports of emotional experience. *Journal of Personality and Social Psychology*, 87(5), 684-697.

Chi, M.T.H. (1997). Quantifying qualitative analyses of verbal data: A practical guide. *Journal of the Learning Sciences*, 6(3), 271-315.

Costa, J.M., & Miranda, G.L. (2017). Relation between Alice software and programming learning: A systematic review of the literature and meta-analysis. *British Journal of Educational Technology*, 48,1464-1474.

Craig, S., Graesser, A., Sullins, J., & Gholson, B. (2004). Affect and learning: An exploratory look into the role of affect in learning with AutoTutor. *Journal of Educational Media,* 29(3)*,* 241-250.

Darabi, A., Arrington, T.L., & Sayilir, E. (2018). Learning from failure: A meta-analysis of the empirical studies. *Education Technology Research and Development,* 66**,** 1101-1118.

D'Mello, S. K., Graesser, A. C., Lehman, B., & Pekrun, R. (2014). Confusion can be beneficial for learning. *Learning and Instruction*, 29, 153-170.

D'Mello, S. K., & Graesser, A. C. (2014). Confusion. *Educational psychology handbook series. International handbook of emotions in education,* 289–310.

Falout, J., Elwood, J., & Hood, M. (2009). Demotivation: Affective states and learning outcomes. *System,* 37(3), 403-417.

Jarosz, A. F., & Wiley, J. (2014). What are the odds? A practical guide to computing and reporting Bayes factors. *The Journal of Problem Solving*, 7(1).

Jordan, M.E., Cheng, AC. J., Schallert, D., Song, K., Lee, SA., & Park, Y. (2014). I guess my question is: What is the co-occurrence of uncertainty and learning in computer-mediated discourse? *International Journal of Computer-Supported Collaborative Learning,* 9(45), 451-475.

Kapur, M. (2008). Productive failure. *Cognition, and Instruction*, 26(3), 379- 424.

Kapur, M. (2012). Productive failure in learning the concept of variance. *Instructional Science,* 40(4), 651-672.

Kapur, M., & Bielaczyc, K. (2012). Designing for productive failure. *Journal of the Learning Sciences*, 21(1), 45-83.

Kapur, M., & Kinzer, C.K. (2009). Productive failure in CSCL groups. *Computer Supported Learning,* 4, 21-46.

Kim, C., & Pekrun, R. (2014). Emotions and motivation in learning and performance. *Handbook of Research on Educational Communications and Technology,* 65-75.

Kort B., Reilly, R., & Picard, R. W. (2001). External representation of learning process and domain knowledge: Affective state as a determinate of its structure and function. In *Proceedings of the AIED Workshops,* 64-69.

Lamnina, M., & Chase, C.C. (2019). Developing a thirst for knowledge: How uncertainty in the classroom influences curiosity, affect, learning, and transfer. *Contemporary Educational Psychology, 59*.

Liu, Z., Pataranutaporn, V., Ocumpaugh, J., & Baker, R. (2013). Sequences of Frustration and Confusion, and Learning. In *Proceedings of Educational Data Mining*, 114-120.

Loibl, K., & Rummel, N. (2014). Knowing what you don't know makes failure productive. *Learning and Instruction,* 34, 74-85.

MacIntyre, P., & Vincze, L. (2017). Positive and negative emotions underlie motivation for L2 learning. *Studies in Second Language Learning and Teaching,* 7(1), 61-88.

McHugh M.L. (2012). Interrater reliability: The kappa statistic. *Biochemia Medica (Zagreb)*, 22(3), 276-282.

Muldner, K., Burleson, B., & VanLehn, K. (2010). "Yes!": Using tutor and sensor data to predict moments of delight during instructional activities. In *Proceedings of the User Modeling, Adaptation and Personalization Conference*, 159-170.

Muldner K., Wixon M., Rai D., Burleson W., Woolf B., & Arroyo I. (2015). Exploring the impact of a learning dashboard on student affect. In *Proceedings* of *Artificial Intelligence in Education Conference*, 307-317.

Nummenmaa, L., & Niemi, P. (2004). Inducing affective states with success-failure manipulations: A meta-analysis. *Emotion,* 4(2), 207-214.

Ozcelik, E., Cagiltay, N. E., & Ozcelik, N. S., (2013). The effect of uncertainty on learning in game-like environments. *Computers & Education*, 67, 12-20.

Pekrun, R., Elliot, A. J., & Maier, M. A. (2009). Achievement goals and achievement emotions: Testing a model of their joint relations with academic performance. *Journal of Educational Psychology,* 101(1), 115–135.

Pietarinen, T., Vauras, M., Laakkonen, E., Kinnunen, R., & Volet, S. (2018). High school students' perceptions of affect and collaboration during virtual science inquiry learning. *Journal of Computer Assisted Learning*, 35, 334–348.

Rittle-Johnson, B., Siegler, R. S., & Alibali, M. W. (2001). Developing conceptual understanding and procedural skill in mathematics: An iterative process. *Journal of Educational Psychology,* 93(2), 346–362.

Tawfik, A.A, Rong, H. & Choi, I. (2015). Failing to learn: Towards a unified design approach for failure-based learning. *Educational Technology Research and Development,* 63, 975-994.

VanLehn, K. (1988). Toward a theory of impasse-driven learning. In *Mandl, D. H. & Lesgold, D. A. (Eds.) Learning Issues for Intelligent Tutoring Systems*, 19–41