Affect and Experience: Case Studies in Games and Test-Taking

Blair Lehman
Educational Testing Service
Princeton, NJ 08540, USA
blehman@ets.org

G. Tanner Jackson
Educational Testing Service
Princeton, NJ 08540, USA
gtjackson@ets.org

Delano Hebert
Educational Testing Service
Princeton, NJ 08540, USA
dhebert@ets.org

Lindsay Grace
American University
Washington, DC 20016, USA
grace@american.edu

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from Permissions@acm.org.
CHI'17 Extended Abstracts, May 06 - 11, 2017, Denver, CO, USA
Copyright is held by the owner/author(s). Publication rights licensed to ACM.
ACM 978-1-4503-4656-6/17/05…$15.00
DOI: http://dx.doi.org/10.1145/3027063.3053341

Abstract
The format of conventional assessments (e.g., multiple-choice) can often make it difficult for students to be motivated or engaged, which can also make it difficult for students to perform to the best of their ability. Game-based assessments (GBA) are a new type of assessment that has been designed to address this issue. We conducted an in-depth investigation of four case studies of user experience during a GBA. The GBA was manipulated in order to investigate whether certain game features (length, feedback) impacted the user experience. The case studies revealed unique patterns in user experience for moment-to-moment changes in enjoyment and frustration. These patterns suggest that GBAs could be utilized to create both a more positive test-taking experience and offer the opportunity for students to perform to the best of their ability.

Author Keywords
User experience; game-based assessment; emotion.

ACM Classification Keywords
H.1.2 [User/Machine Systems]: Human Factors; K.8.0 [General]: Games.

Introduction
Low student engagement and motivation when completing assessments are problematic for the

accurate assessment of student abilities (e.g., [2]). Recent development efforts have focused on designing assessments with a more engaging interaction, which would hopefully motivate students to perform their best. Game-based assessments (GBA) are an example of one of these new assessments [11]. GBAs are proposed to be more engaging because they include game features that have been shown to be engaging in both entertainment [4-6] and serious games [1, 7]. However, there has been little in-depth research of user experience with GBAs. The present research serves as a first step to address this gap in both research and development by conducting an in-depth investigation of four students' user experience with a GBA.

Our investigation of user experience with GBAs was conducted under the framework of the Control-Value Theory [8]. Control-Value Theory posits both what emotions are likely to occur and how those emotions are related to performance by combining expectancy-value approaches, attributional theory, perceived control theory, and models of how emotions impact learning and achievement. Students' perceived level of control over outcomes and value of those outcomes are the determining factors of what emotions will occur. The focus of the present research is on user experience, so we focus on a set of activity emotions (boredom, confusion, engagement, enjoyment, frustration) that Control-Value Theory proposes will occur during task completion and have also been found to occur in game-based learning environments [1, 7].

Control-Value Theory [8] also hypothesizes how emotions will be related to performance. The relationship between emotions and performance is dictated by the valence (positive, negative) and activation (high, low) of the emotion. Positive-activating emotions (enjoyment, engagement) are expected to positively relate to performance, whereas negative-deactivating emotions (boredom) are expected to be negatively related. However, the relationship with negative-activating emotions (confusion, frustration) is not as straightforward. This type of emotion is hypothesized to vary in its relationship to performance based on the frequency and intensity of the emotion as well as student characteristics. Given this potentially complex relationship, we have adopted a detailed case study approach in order to understand user experience with GBAs. Next, we discuss the details of our investigation.

Method

Participants

Participants were 48 students in 6th-, 7th-, or 8th-grade in the northeastern US. The study was completed in a research laboratory and participants received $50 for their participation. In the present work we focus on four students from the larger sample.

Game-Based Assessment

Students completed Text Persuasion, a GBA that was designed to assess students' appeal building skills in a more engaging format than the traditional, multiple-choice format. Figure 1 shows the multiple screens that a student interacts with in Text Persuasion. Students are asked to convince five friends, who are non-player characters, to come to a party in a text-messaging format (see upper left panel of Figure 1). Thus, we are reframing the traditional format (multiple-choice) with a more engaging interface that simulates a familiar activity (text-messaging). In each case the friends present several reasons that they cannot attend the

party and the student must choose one of three text message options (multiple-choice) to convince the friends that they should attend the party (see lower left panel of Figure 1). For example, one friend says that he has to finish his homework. The student then chooses between three responses: (a) offer to help complete it before the party, (b) mention that another friend has already completed it, or (c) state that the friend shouldn't worry about whether or not it is completed.

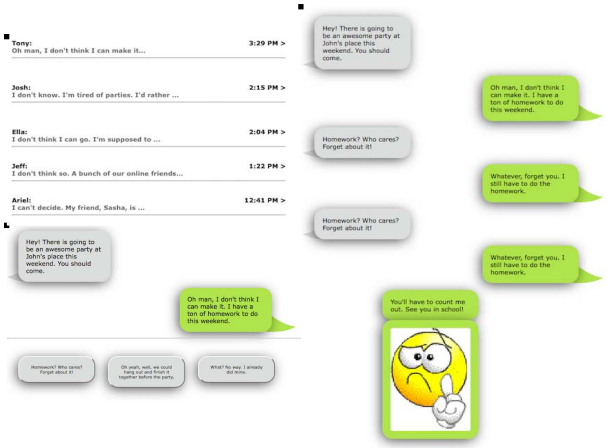

Figure 1: Screenshot of Text Persuasion

Students can either fail (not convince to attend the party) or succeed (convince to attend the party) at each of the five conversations. Failure occurs when students select two responses that are not appealing to the friend, whereas success occurs when students are able to provide an argument that appeals to each obstacle for attending the party by selecting the pre-defined correct response option. At the end of the conversation the friends state whether or not they will attend the party with an emoji that reinforces the failure or success of the conversation (see right panel of Figure 1 for a failed conversation). Each friend has six reasons that he or she cannot attend the party.

In the larger study we manipulated two game features: feedback (present, absent) and game length (long, short). The previously described version of Text Persuasion was the long game length with feedback version. The feedback absent condition involved removing the statement of whether or not the friend would attend and the emoji. The short game length condition involved conversations in which the friends had three reasons that they could not attend the party.

Conventional Assessment

The conventional assessment (CA) was designed to assess appeal building in the Argumentation Learning Progression [10] with 14 multiple-choice items.

User Experience

Students were asked to report on their user experience in two ways. First, students were asked to rate their overall levels of boredom, confusion, engagement, enjoyment, and frustration on a six-point scale (very low to very high) after playing Text Persuasion.

Second, after completing the GBA, students were asked to complete a retrospective affect judgment task that involved watching a replay of synchronized videos of their face and screen from the interaction with Text Persuasion (see Figure 2). While watching these videos, students were asked to indicate whenever there were changes in their level of enjoyment, frustration, or both using the Enjoyment-Frustration Grid in the right panel of Figure 2. The grid is six by six with ranges of very low to very high intensity for both emotions.

Student Characteristics
Students were also asked to complete a survey to collect background information such as grade, gender, and digital game play frequency. Students also completed the Academic Grit scale [3]. Grit has been found to be related to GPA for adolescents and to time spent on academic tasks, which makes this measure a potential proxy for high achieving students.

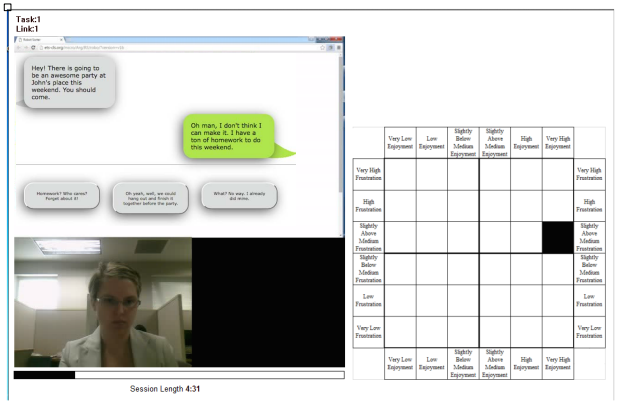

Figure 2: Screenshot of retrospective affect judgment task

Data Processing
To compare user experiences across students, we standardized each student's game playtime by dividing each session into 100 equally sized bins. Thus, all game plays could be compared on the same scale (e.g., 10% or 70% of total playtime) regardless of condition or play length. Student game play actions and emotion ratings were then plotted on the same graph, with time on the x-axis and emotion intensity on the y-axis.

Figures 3 through 6 display the four case studies that will be discussed. In Figure 4, for example, we can see the conversations in which the student engaged in along the top of the graph. Red conversations (Ella and Jeff) represent failures, whereas green conversations represent successes. Similarly, red and green vertical, dotted lines represent failures and successes at the individual decision points. Lastly, the solid blue line represents students' enjoyment intensity and the solid orange line represents students' frustration intensity.

In order to compare performance across the CA and GBA, we divided student performance into three roughly even groups based on low, middle, and high performers for each assessment. We then compared group assignment across the assessments for each student. If the student belonged to the same group in each assessment (e.g., high and high) then there was a match, but if the student belonged to different groups (e.g., high and low) then there was a mismatch.

Results
The results are divided into two sections: (1) performance match and (2) performance mismatch. In each section we discuss the differences in user experience based on students' self-reported emotions.

Assessment Performance Match
The first case was from a female, seventh grade student (Student #1), who performed well on both assessments (High/High). She played digital games infrequently, but had a high level of grit. This suggests that she may find challenges motivating. She was in the short game length with no feedback condition. Overall, she reported an enjoyable (5/6) and engaging (5/6) experience (boredom = 1, confusion = 1, frustration = 1). Next, we discuss her moment-to-moment user experience as shown in Figure 3.

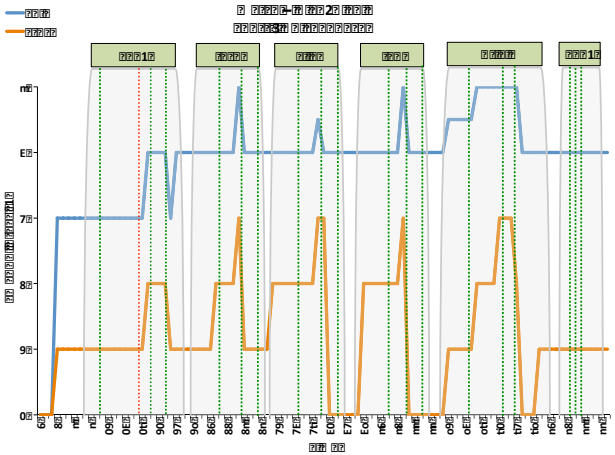

Figure 3: Moment-to-moment user experience for Student #1

Inspection of Figure 3 revealed two patterns. First, this student consistently had a positive experience, as would be expected from her overall ratings. A more positive experience paired with better performance is consistent with Control-Value Theory [8] and empirical findings for emotions during traditional assessments [9]. However, she did have instances of frustration that were higher in intensity than her overall reported experience. This may suggest that she found these momentary challenges motivating rather than discouraging, as would be suggested by her high level of grit. Second, changes in enjoyment and frustration typically occurred at the same time and in the same direction. In other words, increases in enjoyment typically co-occurred with increases in frustration. This student may reflect the ideal case in which the task is sufficiently challenging (increases in frustration), but not to the degree that it creates aversion (consistently higher intensity of enjoyment over frustration). Thus, Student #1 had the balance of enjoyment and frustration that should be desired in GBAs.

The second case was from a female, sixth grade student (Student #2), who did not perform well on either assessment (Low/Low). This student was low on digital game play and grit. Given her low level of construct knowledge, the GBA could be challenging and she may not respond well to it. She was in the long game length with no feedback condition. She reported a slightly more negative user experience than Student #1 (boredom = 2, confusion = 2, frustration = 1, engagement = 4, enjoyment = 5). Next, we discuss her user experience as shown in Figure 4.

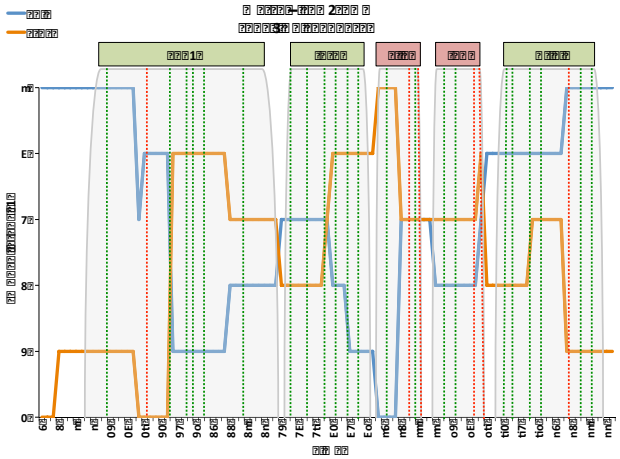

Figure 4: Moment-to-moment user experience for Student #2

A very different pattern emerges when we inspect the moment-to-moment user experience of Student #2 compared to Student #1. First, there are shifts in which emotion is most intense throughout game play. The higher levels of frustration alone would not be predicted

to occur with lower performance based on Control-Value Theory. However, high frustration paired with lower enjoyment, a lower interest in digital games, and low academic grit may be less than ideal conditions for this student to perform well. Second, the changes in enjoyment and frustration intensity typically occurred in opposite directions, such that an increase in enjoyment was paired with a decrease in frustration and vice versa. It may be the case that the task was too challenging based on her skill level. However, there were times when frustration was more intense than enjoyment but she was performing well on the GBA. Perhaps if Student #2 had received feedback about her progress, frustration could have been reduced and she could have had a more enjoyable experience.

Assessment Performance Mismatch

The third case was from a male, eighth grade student (Student #3), who performed well on the CA (High), but not on the GBA (Low). This student played digital games frequently, but had a low grit score. This suggests that the more game-like format could be appealing to him. However, the lower grit score might suggest that he would struggle to persist when not doing well. He was in the short game length with feedback condition. Overall, he reported an experience with more negative (boredom = 2, confusion = 1, frustration = 4) and less positive emotions (engagement = 4, enjoyment = 4) than Student #1, who was also expected to perform well. Next, we discuss his user experience as shown in Figure 5.

Student #3 was expected to perform well on the GBA, so we might expect a similar user experience to Student #1. However, the pattern of user experience for Student #3 was very similar to Student #2, who was expected to and did perform poorly on the GBA. Specifically, both had (a) shifts in which emotion was more intense and (b) changes in emotion intensity that were in the opposite direction. These similarities may suggest that shifts in the more intense emotion and direction of emotion changes are important indicators of user experience and performance on GBAs. Despite playing digital games more frequently, it may be the case that a GBA was not the most appropriate assessment format for this student.

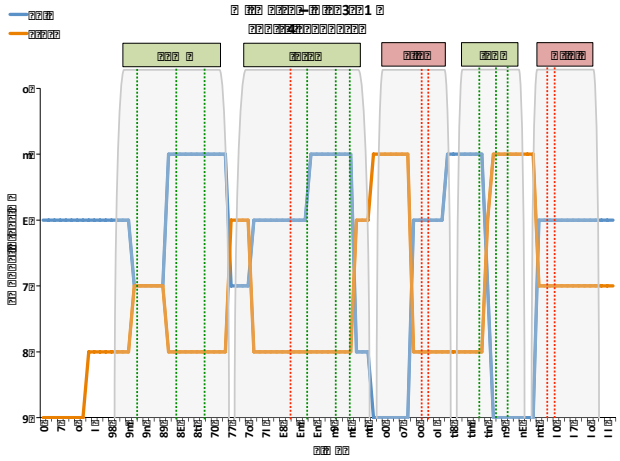

Figure 5: Moment-to-moment user experience for Student #3

The fourth case was from a female, eighth grade student (Student #4), who did not perform well on the CA (Low), but did perform well on the GBA (High). She played digital games frequently and had a high grit score. Student #4 may be a case in which GBAs afford her the opportunity to show her true level of knowledge that may have not been captured by the CA. She was in the short game length with no feedback condition. Overall, she reported a mix of positive (engagement =

4, enjoyment = 4) and negative emotions (boredom = 2, confusion = 3, frustration = 1). Next, we discuss her user experience as shown in Figure 6.

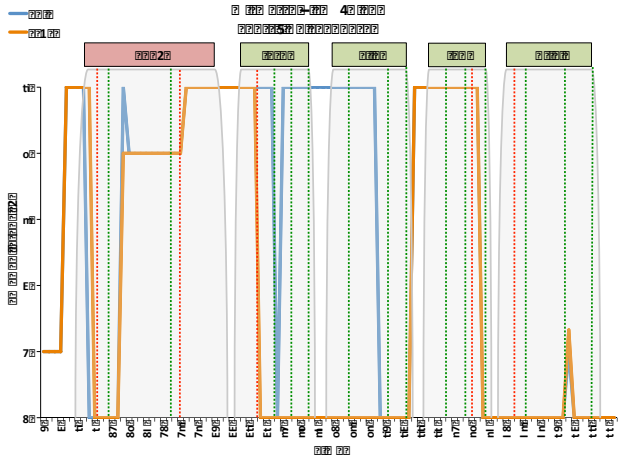

Figure 6: Moment-to-moment user experience for Student #4

The examination of Student #4's interaction reveals a pattern more similar to Student #1 (also performed well) than Student #2 (also expected to perform poorly). Specifically, the changes in emotion intensity between enjoyment and frustration typically occurred in the same direction. However, it is important to note that there are extended periods of time in which she experienced enjoyment and frustration at the same level (instances in Figure 6 when there only appears to be an orange line). Although we proposed for Student #1 that the ideal balance is when changes occur in the same direction but enjoyment is still consistently higher, this does not appear to be the case for Student #4. Control-Value Theory suggests that an emotion such as frustration will be differentially related to outcomes based on student characteristics. Student #4 both played digital games frequently and had a high level of academic grit. Thus, she may have found the extra challenge associated with frustrating experiences motivating, which allowed her to show her true level of knowledge of the construct.

Conclusions

We conducted an in-depth investigation of four user experiences with Text Persuasion, a GBA for appeal building in argumentation. These case studies highlighted three important types of student experiences with a new assessment: students who perform as expected (Students #1 and #2), students who underperform (Student #3), and students who excel (Student #4). This investigation showed that moment-to-moment investigations of students' cognitive and affective user experience can yield useful information about these three types of students. Specifically, the findings from these case studies have provided three unique insights that will facilitate investigation of the larger sample and future research. First, the user experience was generally positive, even when students did not perform well. This suggests that reframing the traditional, multiple-choice assessment with a more engaging interface is promising.

Second, the amount of shifts in which emotion had higher intensity and the direction of emotion changes (same vs. opposite direction) may be important factors for understanding how user experience relates to performance. In particular, we found that how students performed, regardless of expectations, differed on these two factors. We can come to the preliminary conclusion that GBAs should be designed such that moments of challenge and enjoyment co-occur and that

increases in the amount of challenge could be used to inspire students to perform to the best of their ability.

Third, these case studies have provided further support for incorporating student characteristics (e.g., digital game play, academic grit) into analyses. It would not have been possible to find these patterns with only the use of post-game surveys. Thus, the use of methods like the retrospective affect judgment task can provide critical information to understand the user experience, and specifically how different game features can impact the user experience. The combination of these three findings may allow for the determination of how to match students with an assessment format such that students can have both a positive test-taking experience and perform to the best of their abilities.

Acknowledgements

The authors wish to thank Tom Florek and Srinvasa Pillarisetti for developing the recording and retrospective affect judgment technology.

References

1. Ryan Baker, Sidney D'Mello, Ma. Mercedes Rodrigo, and Arthur Graesser. 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. *Int J of Human-Computer Studies*, 68, 4: 223-241.
2. Jürgen Baumert, and Anke Demmrich. 2001. Test motivation in the assessment of student skills: The effects of incentives on motivation and performance. *Eur J of Psych of Ed*, 16: 441–462.
3. Angela Duckworth and Patrick Quinn. 2009. Development and validation of the Short Grit Scale (GritS). *J of Personality Assessment*, 91: 166-174.
4. Kiel Gilleade and Alan Dix. 2004. Using frustration in the design of adaptive videogames. In *Proceedings of the 2004 ACM SIGCHI Conference on Advances in Computer Entertainment Technology*, 228-232.
5. Seung-A Jin. 2012. "Toward integrative models of flow": Effects of performance, skill, challenge, playfulness, and presence on flow in video games. *J of Broadcasting & Electronic Media*, 56, 2: 169-186.
6. Regan Mandryk, and M. Stella Atkins. 2007. A fuzzy physiological approach for continuously modeling emotion during interaction with play technologies. *Int J of Human-Computer Studies*, 65, 4: 329-347.
7. Scott McQuiggan, and James Lester. 2009. Modeling affect expressions and recognition in an interactive learning environment. *Int J of Learning Technology*, 4, 3-4: 216-233.
8. Reinhard Pekrun. 2006. The control-value theory of achievement emotions: Assumptions, corollaries, and implications for educational research and practice. *Ed Psych Rev*, 18, 4: 315-341.
9. Reinhard Pekrun, Thomas Goetz, Anne Frenzel, Petra Barchfield, and Raymond Perry. 2011. Measuring emotions in students' learning and performance: The Achievement Emotions Questionnaire (AEQ). *Cont Ed Psych*, 36, 1: 36-48.
10. Yi Song, Paul Deane, Edith Aurora Graf, and Peter van Rijn. 2013. Using argumentation learning progressions to support teaching and assessments of English language arts. *R&D Connections*, 22: 1-14.
11. Diego Zapata-Rivera and Malcolm Bauer. 2012. Exploring the role of games in educational assessment. In *Technology-based Assessments for 21st Century Skills: Theoretical and Practical Implications for Modern Research*, Michael Mayrath, Jody Clarke-Midura, Daniel Robinson, and Gregory Shaw (Eds.). Information Age Publishing, Inc., Charlotte, NC, 147-169.