How does Students' Affect in Virtual Learning Relate to Their Outcomes? A Systematic Review Challenging the Positive-Negative Dichotomy

Shamya Karumbaiah
University of Pennsylvania
shamya@upenn.edu

Ryan S Baker
University of Pennsylvania
ryanshaunbaker@gmail.com

Yan Tao
University of Pennsylvania
taoyan@upenn.edu

Ziyang Liu
University of Arizona
ziyang773@email.arizona.edu

ABSTRACT

Several emotional theories that inform the design of Virtual Learning Environments (VLEs) categorize affect as either positive or negative. However, the relationship between affect and learning appears to be more complex than that. Despite several empirical investigations in the last fifteen years, including a few that have attempted to complexify the role of affect in students' learning in VLE, there has not been an attempt to synthesize the evidence across them. To bridge this gap, we conducted a systematic review of empirical studies that examined the relationship between student outcomes and the affect that arises during their interaction with a VLE. Our synthesis of results across thirty-nine papers suggests that except engagement, all of the commonly studied affective states (confusion, frustration, and boredom) have mixed relationships with outcomes. We further explored the differences in student demographics and study context to explain the variation in the results. Some of our key findings include poorer learning outcomes arising for confusion in classrooms (versus lab studies), differences in brief versus prolonged confusion and resolved versus persistent confusion, more positive (versus null) results for engagement in learning games, and more significant results for rarer affective states like frustration with automated affect detectors (versus student self-reports). We conclude that more careful attention must be paid to contextual differences in affect's role in student learning. We discuss the implication of this review for VLE design and research.

CCS CONCEPTS

• **Human-centered computing** → Human computer interaction (HCI); HCI theory, concepts and models; • **Social and professional topics** → User characteristics; • **Applied computing** → Education; Interactive learning environments.

KEYWORDS

Affective computing, Student affect, Virtual learning, Student outcomes, Online tutor, Education, Systematic review

ACM Reference Format:
Shamya Karumbaiah, Ryan S Baker, Yan Tao, and Ziyang Liu. 2022. How does Students' Affect in Virtual Learning Relate to Their Outcomes? A Systematic Review Challenging the Positive-Negative Dichotomy. In *LAK22: 12th International Learning Analytics and Knowledge Conference (LAK22), March 21–25, 2022, Online, USA.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3506860.3506863

1 INTRODUCTION

Affective computing in education is an area of research that investigates student affect that arises during learning with the goals of recognizing, measuring, analyzing, and responding meaningfully [1]. Acknowledging the integral link between emotion and cognition, research on affect in Virtual Learning Environments (VLEs) aims to "narrow the communicative gap between the highly emotional human and the emotionally-challenged computer" [2, p.18]). Student affect in intelligent tutoring systems and other types of adaptive and artificially intelligent educational systems has been shown to correlate with a range of other important constructs, including self-efficacy, motivation, and learning [2]. Research has shown that affect plays three primary roles in learning and education: signaling, evaluation, and modulation. These roles refer to the ability of affective states to draw attention to learning challenges, help learners appraise their learning, and help guide cognitive focus [3].

Studies in the past decade have built good quality automated affect detectors in VLEs using physical and physiological sensors, and interaction log data. Additionally, with increasingly available out-of-the-box affect recognition software, it is becoming much easier to augment VLEs with affective capabilities [5], although these technologies do not directly measure the affect most relevant to learning and must be adapted to do so [6]. These detectors have been used in affect-sensitive interventions designed to improve learning gains, and overall experience. For instance, considerable research has investigated the development of affect-sensitive or affect-aware tutoring systems wherein the student experience is personalized by the VLE's ability to detect and respond to students' affective states [2], [4], [5], [19]. The foundational hypothesis of this research is that detecting and responding to student affect improves

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 ACM 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.

LAK22, March 21–25, 2022, Online, USA
© 2022 Association for Computing Machinery.
ACM ISBN 978-1-4503-9573-1/22/03…$15.00
https://doi.org/10.1145/3506860.3506863

the quality of students' interaction with the VLE by making it more engaging and effective for learning. This principal role for affect is justified by the viewpoint that "affective processes are inextricably bound to cognitive and metacognitive processes during learning" [20, p. 2]. Hence, by understanding the relationship between different affective states and student learning in VLEs, these studies aim to design more engaging VLEs that motivate students to learn better.

Several emotion theories in learning incorporate a hedonic principle in which it is assumed that affect is either positive or negative and the goal is for students to experience the positive affective states and avoid the negative affective states [7], [8]. For instance, Pekrun et al. [9] suggest that positive emotions promote learning by improving student motivation and focusing them on the learning activity. As such in some cases, affect is conceptualized using positive and negative valence scales (e.g., PANAS - Positive and Negative Affect Schedule [10], valence-arousal affect grid). Even when affect is conceptualized as discrete states (a more popular choice in VLEs), it is common to assume some states to be inherently positive for learning (e.g., flow) and others to be detrimental to learning (e.g., boredom) [7]. However, the relationship between affect and learning appears to be more complex than this. For example, some studies suggest that positive emotions devoid of self-regulation and motivation may not improve learning [11], while some negative emotions may promote learning by triggering motivation to learn better [12].

A few empirical studies in the past have tried to complexify the role of different affective states while learning with VLEs, beyond the commonly assumed positive-negative dichotomy. First, Liu et al. [17] investigated the overlapping relationships of confusion and frustration as they relate to learning, which was followed by the examination of the possibility of combining these affective states into a joint state referred to as "confrustion" [18], a finding that matches more recent evidence that these two affective states tend to co-occur [30]. Second, certain specific roles that affective states play during student learning have been hypothesized and tested, including Graesser et al.'s [19] investigation of hopeless confusion versus productive confusion and D'Mello et al.'s [20] hypothesis on confusion and frustration arising from logical impasses. Third, some studies have looked at the ways that affect is shaped by specific learning activities (e.g., [21] and [22] studied affect when learning with or without scaffolds). Although it is common to study affective states as they occur in VLEs in general, there has also been an interest in investigating affective-cognitive processes around specific subject matter or educational experiences which are hypothesized to induce stronger emotions. This includes the confrustion surrounding erroneous examples [15], the affective impacts of activities such as medical training scenarios where the patient always dies [24], and writing about traumatic topics [25]. In these contexts, cognition and affect have generally been assumed to be more strongly connected than in everyday learning - making affect instrumental to learning.

In the last fifteen years, several empirical studies in VLEs have investigated the relationship between student affect and their outcomes [13] – [16] (referred from now on also as affect-outcome relationships). However, there has not yet been an attempt to synthesize the results from these empirical studies to more conclusively understand the complex relationship between each affective state and outcome measures of broader importance. Moreover, no study has systematically investigated the contextual factors that complicate the affect-outcome relationship in practice - an aspect that is often overlooked in emotional theories informing VLE design. Hence, to bridge this gap for future learning analytics research and design, we conduct a systematic review of affect-outcome research in VLEs with the following research question in mind - *How does students' affect in virtual learning environments relate to their outcomes*? In doing so, we want to investigate whether the often assumed positive and negative dichotomy in affective states holds true empirically.

2 METHOD

We conducted a systematic literature review to investigate the relationship between students' affective states and learning in VLEs.

2.1 Criteria for Inclusion

Studies had to meet the following requirements to be included in this review:

Discrete Affective States. This review focuses on the relationship between specific affective states and student outcomes. Hence, we included only studies that specified discrete emotions (e.g., confusion, frustration) and excluded studies where affect was instead conceptualized using positive and negative valence scales (e.g., valence and arousal grid without specifying individual affect, PANAS).

External Measure of Outcomes. Given evidence that improved learning within a VLE does not always translate to better student outcomes beyond the system [34], we included only studies that had an external measure of student outcome (e.g., a post-test of knowledge, standardized test, GPA, later life outcomes) and excluded studies that only reported measures internal to the VLE such as in-system outcomes (e.g., rewards earned in a learning game, number of units mastered) or in-system behaviors (e.g., hint seeking, note-taking). Studies that measured outcomes using only self-reports of learning or other subjective constructs like self-efficacy and interest were also excluded.

Study of Affect-Outcome Relationship. We included only studies that explicitly studied the relationship between affective states and student outcomes, excluding studies that only investigated other properties of affect such as its persistence, co-occurrence, and relationship with other affective states. Similarly, we excluded studies that built a predictor of outcomes from affect without studying specific affective states' relationship to student outcomes. We also excluded studies that described an affect-aware VLE or an affective intervention without explicitly reporting results on the relationships between specific affective states and student outcomes.

Empirical Studies. We included only papers that conducted an empirical study and excluded purely theoretical or conceptual papers, literature reviews, meta-analyses, and training manuals, since the purpose of this systematic review is to synthesize empirical results from the literature on affect-outcome relationships.

Studies in VLEs. We included only studies conducted in VLEs and excluded studies conducted in a learning setting without a digital interface that students interacted with and learned from. Although affect research in non-VLE spaces is equally important, in line with the research goals discussed above, this systematic review is focused only on students' affective experience when interacting with a VLE.

Paper Availability and Language. We included only studies in languages readable by the authors (English, Chinese, Japanese, Portuguese, Spanish, Italian) and excluded papers in languages other than these. We also excluded papers whose full texts were not available after searching in digital libraries that were accessible to the co-authors (ACM Digital Library, ProQuest, EBSCO, JSTOR, and Google Scholar) and requesting access from our university librarian.

2.2 Selection Procedure

The earliest work that explored the relationship between affect and learning in a VLE was searched for in ACM Digital Library and Google Scholar using several keywords including "affect", "affective states", "tutoring system", "virtual learning", and "learning". The earliest work found was Craig, Graesser, Sullins, and Gholson's [2] paper titled "Affect and learning: An exploratory look into the role of affect in learning with AutoTutor". This paper is generally known to be seminal in the field, and set out several key aspects of work that followed it (including both methods and the choice of affective states studied). The literature search traversed this paper's entire citation tree, with that paper at its root as of 09/17/2020. Fifty-six papers were excluded for being inaccessible or in a language none of the authors could read (see the last inclusion criteria). Out of the total 1861 unique papers searched, thirty-nine papers were included in the systematic review.

2.3 Coding Reliability

To measure the reliability of inclusion-exclusion coding, 186 papers (10%) were arbitrarily selected to be coded by two independent coders. The two coders were trained in a one-hour session with a coding manual detailing the inclusion-exclusion criteria. The interrater reliability calculated by Cohen's Kappa was 0.869, and the percentage of agreement was 98.92%. Disagreements were resolved through discussion among the two coders to reach a full consensus.

3 FINDINGS

3.1 Overview

The studies varied in their measures of student outcomes. The majority of the studies (29 out of 39) were conducted over a limited time, and student outcomes were measured using some sort of knowledge test that was administered at the beginning and end of the study. These studies mainly focused on subject matter learning in areas such as math (6), languages (4), history (3), computer programming (3), and computer literacy (3). There were four studies that measured student outcomes in academic courses using longer-term measures like exam scores and GPA. This includes a math course, two programming courses, and one computer literacy course. There were two studies that measured student outcomes using standardized test scores - both focused on math learning. And finally, there were four studies that measured longer-term career outcomes including college enrollment and STEM major choice. The studies also varied vastly in duration. Laboratory studies with short-term outcomes were less than an hour long, while classroom studies with both short and long-term outcomes varied from several hours to several years. The implications of this difference will be discussed further later. The majority of the studies were conducted in the United States (21), more than entire continents (Asia 9; Europe 7; other North American countries 2). The majority of the studies were conducted with undergraduate students (26), followed by middle school (9), high school (4), graduate (3), and elementary school students.

3.2 Commonly Studies Learning-Centered Emotions

Among the 39 studies reviewed in this paper, the four most commonly studied affective states are engagement/engaged concentration/flow (26 studies), confusion (20 studies), frustration (20 studies), and boredom (22 studies). Despite the different terms, engagement, engaged concentration, and flow are typically used to indicate the same affective state (see discussion in [20]). Here, we present a summary of how each of these affective states relates to student outcomes by synthesizing the results from all papers that reported it. We analyzed the data using the counting method (number of studies positive, negative, and null; [31]), random effects models, and three-level models [33]. The overall results of the three methods were identical. Within this paper, we report the counting method because it allows for greater exploration of the degree to which a finding is consistent across studies.

3.2.1 Generally Positive Relationship of Engagement with Learning Outcomes.

In the 26 studies that examined the affective state of engagement, it was either observed to be positively related to the learning outcome or had a null relationship. None of the 26 studies reported a negative relationship. In eleven studies the relationship was positive [13], [14], [21], [27], [35] – [41] and in seven it was null [19], [32], [42]–[46]. In contrast, eight studies reported mixed results (null and positive) within the same study. This was a result of different choices and contexts in these studies, including:

- different statistical approaches used to measure the relationship (e.g., positive when affect was taken by itself and null in more complete models where collinearity was ignored [29], [2], [47])
- varied outcome measures in the same study (e.g., positive for college enrollment but null for STEM major choice in [48], positive for an immediate knowledge test and null for a delayed test in [49])
- the VLE behavior changed over time (e.g., null when the VLE was scaffolding using hints and explanations and positive when the scaffolds faded out later, in [22])
- different participant population (e.g., positive with undergraduate students but null when tested for replicability in a crowdsourcing platform (Amazon Mechanical Turk) in [25])

Engagement was studied in seven learning games, where its relationship with student outcomes is of particular interest, given the focus on engagement in game design [50]. In the majority of the studies conducted in learning games (5 out of 7), engagement was related positively to student learning – in history [38], social studies [36], language [35], [39], and complex processes [40]. In the other two studies, engagement did not have a significant relationship with student learning (e.g., social study learning in [45] and environmental protection in [46]). Brom et al. [40], [49] differentiated flow as an affective state from engagement as a cognitive state ("learning involvement"). However, the relationships with outcome were the same for flow and engagement in both the studies.

3.2.2 Strong Disagreement on How Confusion Relates to Learning Outcomes. Unlike engagement, the relationship between confusion and student outcomes varied considerably among the 20 studies that reported it. The distribution is as follows:

- five positive [2], [13], [19], [42], [47]
- seven null [20], [25], [26], [32], [43], [44], [51]
- three both positive and negative - the results within a study varied based on the version of the VLE (e.g., positive with scaffolds and negative without in [21]), the statistical method used to analyze the data (e.g., positive when affect was taken by itself and negative in more complete models where collinearity was ignored [29]), and the persistence of confusion (positive for resolved confusion and negative with prolonged confusion [23])
- three both negative and null - the results within a study varied based on the VLE version (null with a conversational agent and negative without in [14], negative with scaffolds and null without in [22]), and the learning outcome (negative for college enrollment and null for STEM major choice in [48])
- two negative [27], [37]

Some studies also highlighted the nuanced relationship of confusion with learning by exploring not only the incidence of confusion but also its persistence. For instance, both Gong et al. [47] and Lee et al. [23] observed a positive association between resolved confusion and learning, while Lee et al. [23] also observed a negative association between prolonged confusion and learning. Lee et al. [23] operationalized resolved confusion as confusion followed by at least two 20-second clips of non-confusion and prolonged confusion as persistent confusion for at least three 20-second clips. Similarly, Rodrigo et al. [14] reported that confusion had a positive relationship with learning when it was preceded or followed by engagement.

There were also significant differences in how confusion was measured - either by how it was expressed (e.g., puzzled facial expression or gesture) or using cognitive aspects of student experience (e.g., struggling with learning). Across the 20 papers that reported results for confusion, some studies focused more on the expression [14], [27], [44], while others on the cognitive aspects (e.g., Lee et al. [23] coded confusion when students made repeated errors). As confusion is often theorized to be an important affective state for student learning [8], in a later section we further contextualize the differences in these studies in terms of geographic location, affect collection method, and authenticity of the research setting to better understand the high degree of variance in the results for confusion.

3.2.3 A Few Negative but Mostly Non-significant Results for Frustration. The majority of the papers that studied the relationship between frustration and student outcomes did not report a significant result (13 out of 20: [2], [14], [19], [22], [25], [27], [28], [37], [42]–[44], [43], [52]). Some of them attributed this to the observed low incidence of frustration during student learning. Of the seven studies that reported a significant result for frustration, two reported a negative relationship [5], [26], and four reported a combination of negative and null relationships. The mixed results (both negative and null) were observed when:

- frustration was reported at different phases of learning. D'Mello et al. [20] observed a negative relationship when students were asked to report frustration halfway through the current problem or right after (3 seconds after) receiving feedback on the previous problem. However, this result stopped being significant when students reported frustration at the onset of the current problem (7 seconds after the current problem was displayed) or when students spontaneously reported affect at an arbitrary time.
- the outcome measure varied (e.g., negative for college enrollment and null for high-school course choice (both AP math and science) and STEM major choice in [32], [48])
- different statistical approaches used to measure the relationship (e.g., negative with correlation analysis but null with mediation analysis in [51])

Only one study reported a positive relationship and noted it to be an unexpected result [21].

Although frustration was associated negatively with learning in [26], it was observed that frustration preceded or followed by confusion was positively associated. A small subset of studies has tried to merge the affective states of confusion and frustration into a single affective state (called confrustion) due to their overlapping conceptualization, especially when measured using cognitive aspects of student experience (e.g., repeated errors while learning). Richey et al. [15], [18] studied confrustion and observed that it was negatively associated with learning. In contrast, Liu et al. [17] observed that brief episodes of confusion and frustration when analyzed together were associated positively with learning gain in math learning.

3.2.4 Mostly Negative or Null Relationship Between Boredom and Learning Outcomes. Among the 22 studies that reported the relationship between boredom and learning outcomes, the results were mostly negative or null:

- four negative [26], [27], [37], [47]
- nine null [14], [19], [20], [25], [42]–[44], [52], [53]
- four both negative and null (e.g., negative for college enrollment and null for STEM major choice in [48], negative in correlation analysis, null in more complete models where collinearity was ignored [2], [51], negative for low-pretest students and null for high-pretest students in [54])
- two negative, null, and positive (based on whether affect was taken by itself or in more complete models where collinearity was ignored [29] and based on the outcome measure in the same study - positive for college enrollment, null for high-school course choice (AP math), negative for high-school course choice (AP science) in [32])
- two both negative and positive (e.g., negative when not scaffolded and positive when scaffolded with hints and explanations - observed in a math VLE [21] and in a computer programming VLE [22])
- one positive [13]

Table 1: Results for other affective states that were reported in at least five studies

Affect	#Null Results	#Non-Null Results	Result[a]	Studies
Surprise	7	1	0	[19], [20], [25], [27], [42], [44], [51]
			+/0	[47]
Delight	5	0	0	[19], [27], [42], [44], [47]
Enjoyment	4	2	+	[40], [49], [52], [53]
			0	[55] (medical education),[51] (math learning)
Neutral	8	1	0	[2], [19], [20], [22], [25], [27], [42], [44]
			-/0	[47]

[a] 0 null, + positive, +/0 positive or null, -/0 negative or null

Despite having relatively more null results than negative results, boredom was often concluded to be detrimental for learning. For instance, after hypothesizing boredom as a "direct antithesis" to Csikszentmihalyi's [7] "zone of flow", Craig et al. [2] concluded a negative link between boredom and learning even when the negative relationship was only observed in the correlational analysis and not in the regression and ANOVA analyses. In the few papers that did report a positive result for boredom, it was often interpreted as an unexpected result with a potential explanation - usually attributing it to high performing students getting bored. For instance, when Bosch et al. [22] observed a positive relationship between boredom and student performance when scaffolding (with hints and explanations), boredom was attributed to scaffolds being less challenging. Similarly, in the causal modeling conducted by Fancsali [13], it was suggested that a missing unmeasured variable (e.g., scaffolding) was potentially causing boredom in high-performing students. However, there is not yet empirical evidence for these possible accounts for the complex relationship that sometimes emerges between boredom and learning outcomes.

3.3 Mostly Null Relationship for Other Affective States

Anxiety was investigated in four studies. Its relationship to learning was reported as negative in two studies (both language learning - [39] and [41]) and null in two studies [20], [51]. Beyond anxiety, the majority of the reported results for the less commonly-studied affective states were null results, including surprise, delight, enjoyment, and neutral. Table 1 lists the affective states for which results were reported in at least 5 studies. Several papers studied affective states other than just engagement, confusion, frustration, and boredom, but the majority of the studies did not include these affective states in the analysis due to their very low occurrence preventing the authors from running quantitative analysis. It is worth noting that the lack of findings for these affective states may be a reflection of the limitations of the methodologies currently used in this research. For example, the rare or brief occurrence of a strong affective state (e.g., anxiety, eureka) may turn out to be more important than is currently realized. The theoretical frameworks in Kort et al. [8], Craig et al. [2] hypothesized that eureka is associated with students acquiring profound new insights during learning and thus, linked to better learning. But in the 20 hours of their study, Craig and colleagues [2] only observed one instance of eureka. The authors suggested that the observation duration for each student (30 seconds every 5 minutes) was probably too far apart to capture the rapid experience of a state like eureka.

Table 2: Counts of the reported relationship (+, NULL, -) between the different affective states and learning outcomes aggregated based on the three continents[a]

Affect		ENG			CON			FRU			BOR		
Relationship	N	+	∅	-	+	∅	-	+	∅	-	+	∅	-
North America	23	9	10	0	6	7	5	2	11	5	6	11	9
Asia	9	6	3	0	2	2	3	0	4	0	0	2	2
Europe	7	4	2	0	0	1	0	0	2	1	0	2	1

[a] Notable differences highlighted in grey

3.4 Differences in Affect-Outcome Relationship Based on Contextual Factors

In this section, we consider whether the results seen could have been influenced by student demographics (geographic location) and study context (learning environment, affect data collection protocol). The goal is to explore the differences in the results when the studies are categorized based on these contextual factors.

3.4.1 Differences in Geographic Location – Cultural or Methodological? Differences in culture influence variation in beliefs and personal dispositions towards emotional expression and moderation and the frequency and emergence of certain affective states [56]. Due to a limited number of studies at the country-level (with an exception of the United States; see Section 3.1), we categorized studies at the continent-level (Table 2). Accordingly, the majority of the studies were conducted in North America (~60%), followed by Asia (~23%) and Europe (~17%). Affect-Outcome studies in North America investigated all the four learning-centered emotions more often than Asia (fewer studies investigated frustration and boredom) or Europe (results reported mainly for engagement).

Engagement. The results for the relationship between engagement and student outcomes are split almost equally between positive (9) and null (10) in North America. In contrast, there is more positive than null in Asia (6 vs 3) and Europe (4 vs 2) - double in both cases. This may be because of differences in the VLEs studied

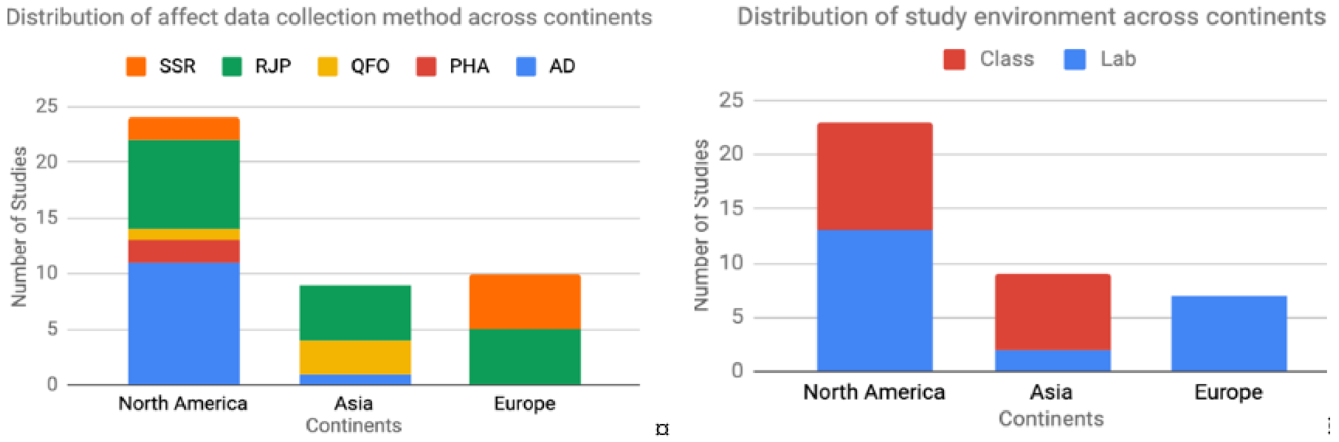

Figure 1: The distribution of affect data collection method (left) and study environment (right) across North America, Asia, and Europe

in different continents. All but two studies from Europe investigated student affect in learning games, and four out of nine studies from Asia involving engagement were also conducted on a learning game. In comparison, none of the studies from North America involving engagement used a learning game. As elaborated earlier, engagement was positively related to learning outcomes in the majority of the studies conducted in learning games. Hence, this difference could possibly be attributed to the differences in types of learning experiences as opposed to the geographic location.

Confusion. The affect-outcome relationship for confusion appears to be more negative in Asia (3 out of 7) as compared to North America (5 out of 18), but a greater number of studies would be needed to consider this conclusive. No other obvious factors appeared to explain this possible difference.

Frustration. All the studies from Asia reported a null relationship between frustration and student outcomes. In contrast, while the majority of results point at a null relationship for frustration in North America (11), there are also a few negative relationships (5) that cannot be ignored, especially since frustration can be easily missed due to its low occurrence. Further research may be necessary to understand the cultural differences in these contexts with respect to the frequency, expressivity, and moderation of frustration.

Boredom. All the studies from Asia and Europe with results for boredom (although few in number) reported a negative or null relationship with student outcomes. All the positive results were reported by studies from North America. However, in most of these studies, a positive result for boredom was interpreted as being unexpected with a possible confounding variable (e.g., presence of scaffolds). The result was usually attributed to high-performing students being bored with less challenging tasks. Further empirical evidence to support these claims and further investigations on the cultural factors that influence this phenomenon may help to better understand the role of boredom in student learning.

Along with cultural factors, another potential explanation for the differences in the results based on the geographic location could come from the differences in the methodological practices associated with the studies in the three continents. As discussed earlier, some of the differences in the results for engagement could potentially be attributed to the type of VLE used (e.g., learning games). To further investigate this, we explore two other study context-related categories (affect data collection method, study environment) that have a high variance across the continents (Figure 1). For instance, all the studies in Europe were conducted in a lab study, while the majority of studies in Asia were conducted in a traditional classroom, and studies in North America were split between the two (Figure 1; right). These categories are explained in detail in their respective sections below.

3.4.2 Types of Measurement of Affect Associated with Significant Differences in Affect-Outcome Results. Affect labeling is inherently subjective and the choice made by the research study design in terms of who will provide labels (student vs outside observer vs algorithm) and when (in real-time vs retrospectively) may have an impact on the data collected, and in turn the relationships found in that data. For instance, past studies have reported issues with the reliability of self-reports [57]. On the other hand, cultural differences between the annotators and students have been reported to interfere with the quality of affect annotations [58]. Hence, we present results for the differences in affect and outcome relationships based on the affect data collection protocol. The affect data were collected using the following 5 techniques:

- Student Self-Reports (SSR) in which students reported their own affect in real-time by answering an affect survey question that popped up in the VLE during their learning.
- Retrospective Affect Judgement Protocol (RJP) in which students reported their own affect but retrospectively after the study or learning session, while watching a playback of their activities during the session (e.g., a facial video captured by a webcam, screen capture of their VLE interaction)

Table 3: Counts of the reported relationship (+, NULL, -) between the different affective states and learning outcomes aggregated based on the affect data collection method used

Affect		ENG			CON			FRU			BOR		
Relationship	N	+	∅	-	+	∅	-	+	∅	-	+	∅	-
SSR	7	5	4	0	0	1	0	0	2	0	0	2	0
RJP	18	10	7	0	3	5	1	1	7	3	2	6	4
QFO	4	3	2	0	1	2	2	0	4	0	0	3	2
PHA	2	0	0	0	0	0	0	0	1	0	0	1	1
AD	12	5	4	0	4	3	5	1	4	3	4	4	5

[a] Notable differences highlighted in grey

- Quantitative Field Observation (QFO) in which a trained coder is present in the classroom or lab making real-time observations of students' affect
- Post-Hoc Human Annotation (PHA) in which a trained coder annotates the digital traces of the student activity (e.g., a facial video captured by a webcam, VLE interaction data)
- Automated Detection (AD) in which a predictive model is built using the previously collected student affect data (e.g., QFO, PHA) to automatically assign affect labels to the digital traces of student activity (e.g., VLE interaction data, facial video), and the predictive model is used in further analysis

As presented in Table 3, the most popular approach to collect affect data in the affect-outcome literature is RJP (18) followed by AD (12). There were four studies that used both SSR and RJP and they were counted for both methods. SSR was used mostly in learning games (5 out of 7). Since these studies mostly focused primarily on engagement (flow), there are fewer results for confusion, frustration, and boredom with SSR. Although human annotation (QFO and PHA) studies are fewer in number (6 in total), it is important to note that human annotations were one of the primary sources of data to build predictive models of affect, which was then applied on a larger dataset to get finer-grained affect labels (AD).

Engagement. There is no notable difference between the approaches in the relationships for engagement.

Confusion. There are relatively more negative results for confusion with AD (5 out of 12) and QFO (2 out of 5) than RJP (1 out of 9). In comparison, RJP has more null results for confusion (5 out of 9).

Frustration. Similar to confusion, there are relatively more negative results for frustration with AD (3 out 8) than RJP (3 out of 11). RJP has a majority of null results (7 out of 11). All the studies with the other methods (SSR, PHA, QFO) only reported null results for frustration.

Boredom. There are relatively more studies with significant results for boredom with AD (9 out of 13) than RJP (6 out of 12). All the studies with the other methods (SSR, PHA, QFO) only reported negative or null results for boredom. Although fewer in number, the majority of the positive results (4 out of 6) for boredom were reported by studies that used AD.

The more frequent significant results with AD and the more frequent null results with RJP as compared to all other methods, especially for relatively rare affective states such as frustration and boredom, raise scientific and methodological questions. First, using a predictive model of affect (AD) enables researchers to obtain finer-grained data for a much larger sample. Sample size alone (both in terms of number of students and number of observations per student) may explain the higher frequency of significant effects for this method. Obtaining a large sample is less feasible with the other methods, all of which require considerable human labor. Second, RJP had the majority of non-significant results, even beyond what might be expected due to its frequency. This raises questions on the reliability of self-reports from students, especially in terms of students' ability to recall discrete emotions they experienced during their learning when asked to reflect on them after the completion of the learning activity. Recent work suggests that there is considerable variation in students' retrospective memory of their own affect, even shortly thereafter [59].

Table 4: Counts of the reported relationship (+, NULL, -) between the different affective states and learning outcomes aggregated based on the environment of data collection[a]

Affect		ENG			CON			FRU			BOR		
Relationship	N	+	∅	-	+	∅	-	+	∅	-	+	∅	-
Classroom	17	9	6	0	4	5	7	1	7	2	4	6	6
Lab	22	10	9	0	4	5	1	1	10	4	2	9	6

[a] Notable differences highlighted in grey

3.4.3 Confusion Arising in Classrooms (vs Lab) Related to Poorer Learning Outcomes.

Another potentially important difference between studies is whether the study was conducted in a traditional classroom or a laboratory setting. Student experience could be more authentic in a natural setting like a classroom than in a controlled setting like a laboratory. Accordingly, the results from a classroom setting could be more generalizable to real-world contexts [60]. Hence, it is important to examine whether there are differences in the affect-outcome relationships between studies conducted in a classroom versus a lab.

As presented in Table 4, across the 39 studies reviewed, 17 of them were conducted in a classroom, while the remaining 22 were conducted in a laboratory. There are no notable differences in the results for engagement, frustration, and boredom.

However, almost all of the negative results for confusion (7 out of 8) were reported by studies conducted in a classroom as opposed to a lab. Several factors may explain this pattern. The laboratory studies were all conducted in a controlled environment with a short learning activity, often around 30-45 minutes long, and a short-term learning outcome which was usually measured with a knowledge test at the end of the study. It is likely that the confusion arising in such simple, unauthentic settings may either be resolved quickly or have a low consequence on student outcomes. In comparison, classroom studies investigated student affect data for a longer time period (6 studies lasting 1-5 years, 3 studies lasting multiple months, 3 studies lasting a few days, 3 studies lasting a few hours, and only 2 studies under an hour). Confusion arising in such natural, authentic learning settings could have a serious impact on student outcomes, especially if it is prolonged or left unresolved [17, 23]. In addition, several of the student outcomes in classroom studies were longer-term and high-stakes, including four career-related outcomes such as college enrollment, four course-related outcomes such as GPA, and two standardized tests. Another explanation could be found in the mismatch in student age or school level. All the lab studies recruited either undergraduate (21) and/or graduate students (3) - two studies had both undergraduate and graduate students. It is possible that older students have a better ability to resolve confusion on their own without much external support and possibly even use it as a tool to learn better. Age is also known to influence emotional expressivity and inhibition [61]. If a VLE is designed specifically for undergraduate students, there should be no harm in recruiting undergraduate students for a lab study. However, if the goal of the study is to conduct a more general theoretical or basic research on student affect with potential implications to younger students, a convenience sampling of undergraduate students may be harmful to the generalizability of the study's claim (see discussion in [62]). Further research is needed to fully examine the ecological validity of laboratory settings for this type of research.

4 DISCUSSION AND CONCLUSIONS

The relationship between student affect and learning is often assumed to be either positive or negative (e.g., boredom and frustration are assumed to be bad for learning). However, the empirical evidence appears to tell a more complex story. By synthesizing the affect-outcome research in VLEs, this systematic literature review aimed to complexify the account of how students' outcome measures relate to their affect that arises during their learning in a digital environment. Here is a summary of this review's key findings on affect-outcome relationships, including contextual factors that may explain some of the differences:

- There is a general consensus on the positive relationship between *engagement* (or engaged concentration or flow) and learning outcomes. No study reported a negative result, while some studies reported null results. Most of the null results were reported in studies from North America as compared to Europe and Asia. Further analysis revealed that the majority of the studies from Europe and Asia were conducted on learning games that were specifically designed to maximize flow, while none of the studies in North America used a learning game.
- *Boredom*, which is often perceived as the antithesis of engagement, has a mostly negative or null relationship with outcomes. In the rare reports of positive results for boredom, it was treated as an unexpected result, and often attributed to a confounding variable (e.g., the presence of scaffolding) making learning less challenging for high-performing students.
- Despite being commonly studied, there was little empirical evidence for a significant relationship between *frustration* and outcome measures. Many studies attributed this to the low occurrence of frustration.
- The affective state with the most variation in its relationship to outcome measures was *confusion.* Along with the differences in how confusion was measured (expression versus cognitive aspects), there was also some empirical evidence for the need to differentiate brief versus prolonged confusion and resolved versus persistent confusion. Further analysis revealed that confusion arising in classrooms was related to poorer learning outcomes, but that this pattern did not manifest in lab studies. The lab studies differed from the classroom studies in that they involved shorter learning activities, short-term and low-stakes outcomes, and older participants (undergraduate or graduate students). This raises questions about the ecological validity of the laboratory studies in understanding the role of confusion in student learning.
- The studies that use automated detectors to generate affect labels see more significant results for rarer affective states (e.g., frustration) compared to studies that ask students to self-report their affect retrospectively after the learning session has completed. Studies with predictive models of affect have a much larger sample size and more fine-grained data. This also raises questions about students' ability to recall finer details of the different emotions they experienced in a learning session that has passed.
- A few other affective states were studied across these papers (e.g., surprise, delight, enjoyment) but they mostly had a null relationship with outcome measures. *Anxiety* had a negative relationship to outcome measures in 2 of 4 studies.

Implications for VLE Design and Development. As discussed in the first section, affective states are often treated as being inherently good or bad, both in emotion theories pertaining to learning and in designing affect interventions in VLEs. However, as demonstrated in this systematic review, affect-outcome relationships are not as straightforward as is generally assumed. Except for engagement, all of the affective states had some mixed results. The pattern of results is particularly complex and nuanced in the case of confusion - an affective state commonly used in affect-sensitive VLE design [19]. Differences in results across continents and methodological choices also raise questions on the generalizability of theoretical and empirical research on the affect-outcome relationship. Hence, more careful attention must be paid in future learning analytics research and design in making assumptions about affect's role in student learning.

Implications for Research. There is considerable variation in findings on the role of confusion in student learning. More research is needed to establish how confusion should be conceptualized to be able to measure it effectively. This review also raised specific questions that warrant further research. First, there is a need to investigate how cultural factors, study design, and methodological choices impact the results on affect-outcome relationships. Second, the ecological validity of laboratory studies needs to be more thoroughly considered. Third, current affect data collection methods may need to be improved to more efficiently capture rare affective states. Fourth, the consistent null results for rare affective states could be a limitation of the current methodological approaches that may be failing to recognize the role of sparse but potentially important events. Fifth, there is a possibility of confounding variables (e.g., motivation influencing both affect and learning), which can be investigated by measuring a wider range of variables beyond affect and outcomes.

Limitations and Future Work. In this review, we focus primarily on commonly-measured academic outcomes. To get a full understanding of the role that affect plays in VLEs, future reviews could benefit from considering other aspects of the learning experience such as self-efficacy. In addition, some studies have explored more nuanced patterns of affect including persistence, transitions, and multi-state sequences [17], [23], [26]. For affective states with less conclusive evidence (e.g., confusion, frustration), it may help to synthesize the results from these studies as well. Also, this review is limited to studies conducted using VLEs where students' affective experiences focus on one-to-one human-computer interaction. The affect-outcome relationship in other learning settings such as collaborative learning in a physical classroom may vary due to the presence of peers and teachers (e.g. [22]). Other factors, such as student demographic factors (age/school level, race/ethnicity, gender), outcome category (learning content test, standardized test, academic course, career), and learning context (subject matter, duration), may also play an important role, and will be valuable to examine.

In summary, this study suggests that except engagement, all other commonly-studied affective states (confusion, frustration, and boredom) in VLEs seem to have a mixed relationship with students' learning outcomes – a finding that contradicts the common assumption of affective states being inherently good or bad. We consider how the affect-outcome relationships vary in terms of study context. We conclude that more careful attention must be paid to contextual differences in affect's role in student learning.

REFERENCES

[1] Rafael A. Calvo and Sidney D'Mello. 2010. Affect detection: an interdisciplinary review of models, methods, and their applications. *IEEE Transactions on Affective Computing* 1, 1 (2010), 18–37.

[2] Scotty D. Craig, Arthur C. Graesser, Jeremiah Sullins, and Barry Gholson. 2004. Affect and learning: an exploratory look into the role of affect in learning with AutoTutor. *Journal of Education Media* 29, 3 (2004), 241–250.

[3] Carroll E. Izard. 2010. The many meanings/aspects of emotion: definitions, functions, activation, and regulation. *Emotion Review* 2, 4 (2010), 363–370.

[4] Shamya Karumbaiah, Rafael Lizarralde, Danielle Allessio, Beverly Woolf, Ivon Arroyo, and Naomi Wixon. 2017. Addressing student behavior and affect with empathy and growth mindset. In *Proc. of the 10th International Conference on Educational Data Mining.* 96–103.

[5] Gustavo Padron-Rivera, Cristina Joaquin-Salas, Jose-Luis Patoni-Nieves, and Juan-Carlos Bravo-Perez. 2018. Patterns in poor learning engagement in students while they are solving mathematics exercises in an affective tutoring system related to frustration. In *Proceedings of the 10th Mexican Conference on Pattern Recognition.* 169–177.

[6] Anabil Munshi, Shitanshu Mishra, Ningyu Zhang, Luc Paquette, Jaclyn Ocumpaugh, Ryan Baker, and Gautam Biswas. 2020. Modeling the relationships between basic and achievement emotions in computer-based learning environments. In *Proceedings of the 21st International Conference on Artificial Intelligence in Education.* 411–422.

[7] Mihaly Cziksentmihalyi. 1990. *Flow: The Psychology of Optimal Experience.* Harper & Row, New York, NY.

[8] Barry Kort, Rob Reilly, and Rosalind W Picard. 2001. An affective model of interplay between emotions and learning: reengineering educational pedagogy—building a learning companion. In *Proc. of IEEE Int'l Conference on Advanced Learning Technologies.* 43–46.

[9] Reinhard Pekrun, Stephanie Lichtenfeld, Herbert W. Marsh, Kou Murayama, and Thomas Goetz. 2017. Achievement emotions and academic performance: longitudinal models of reciprocal effects. *Child Development* 88, 5 (2017), 1653–1670.

[10] David Watson, Lee Anna Clark, and Auke Tellegen. 1988. Development and validation of brief measures of positive and negative affect: the PANAS scales. *Journal of Personality and Social Psychology* 54, 6 (1988), 1063-1070.

[11] Carolina Mega, Lucia Ronconi, and Rossana De Beni. 2014. What makes a good student? How emotions, self-regulated learning, and motivation contribute to academic achievement. *Journal of Educational Psychology* 106, 1 (2014), 121-131.

[12] Marcelo Worsley and Paulo Blikstein. 2015. Using learning analytics to study cognitive disequilibrium in a complex learning environment. In *Proceedings of the Fifth International Conference on Learning Analytics And Knowledge.* 426–427.

[13] Stephen E. Fancsali. 2014. Causal discovery with models: behavior, affect, and learning in Cognitive Tutor Algebra. In *Proceedings of the 7th International Conference on Educational Data Mining (EDM 2014).* 28–35.

[14] Mercedes T. Rodrigo, Julieta Nabos, Loyola Heights, and Worcester Ma. 2010. The relationships between sequences of affective states and learner achievement. In *Proceedings of the 18th International Conference on Computers in Education.* 56–60.

[15] J. Elizabeth Richey, Juan Miguel L. Andres-Bray, Michael Mogessie, Richard Scruggs, Juliana M.A.L. Andres, Jon R. Star, Ryan S. Baker, and Bruce M. McLaren. 2019. More confusion and frustration, better learning: the impact of erroneous examples. *Computers and Education* 139 (2019), 173–190.

[16] Amber Chauncey Strain and Sidney K. D'Mello. 2015. Affect regulation during learning: the enhancing effect of cognitive reappraisal. A*pplied Cognitive Psychology* 29 (2015), 1-19.

[17] Zhongxiu Liu, Visit Pataranutaporn, Jaclyn Ocumpaugh, and Ryan S. Baker. 2013. Sequences of frustration and confusion, and learning. In *Proceedings of the 6th International Conference on Educational Data Mining (EDM 2013).* 114–120.

[18] J. Elizabeth Richey, Bruce M. McLaren, Miguel Andres-Bray, Michael Mogessie, Richard Scruggs, Ryan S.J.d. Baker, and Jon Star. 2019. Confrustion in learning from erroneous examples: Does type of prompted self-explanation make a difference. In *Proceedings of the 20th International Conference on Artificial Intelligence in Education (AIED 2019).* 445–457.

[19] Arthur Graesser, Sidney D'Mello, Patrick Chipman, Brandon King, and Bethany McDaniel. 2007. Exploring relationships between affect and learning with AutoTutor. In *Supplementary Proc. of the 13th Int'l Conf on Artificial Intelligence in Education (AIED 2007).* 14–21.

[20] Sidney K. D'Mello, Blair Lehman, and Natalie Person. 2010. Monitoring affect states during effortful problem solving activities. *International Journal of Artificial Intelligence in Education* 20, 4 (2010), 361–389.

[21] Zachary A. Pardos, Ryan S.J.d. Baker, Maria O.C.Z. San Pedro, Sujith M. Gowda, and Supreeth M. Gowda. 2013. Affective states and state tests: investigating how affect throughout the school year predicts end of year learning outcomes. In *Proceedings of the Third International Conference on Learning Analytics and Knowledge.* 117–124.

[22] Nigel Bosch, Sidney D'Mello, and Caitlin Mills. 2013. What emotions do novices experience during their first computer programming learning session?. In *Proceedings of the 16th International Conference on Artificial Intelligence in Education (AIED 2013).* 11–20.

[23] Diane Marie C. Lee, Ma. Mercedes T. Rodrigo, Ryan S.J.d. Baker, Jessica O. Sugay, and Andrei Coronel. 2011. Exploring the relationship between novice programmer confusion and achievement. In *Proceedings of the 4th International Conference on Affective Computing and Intelligent Interaction (ACII 2011).* 175–184.

[24] Jeanine A DeFalco, Jonathan P Rowe, Luc Paquette, Vasiliki Georgoulas-Sherry, Keith Brawner, Bradford W Mott, Ryan S Baker, and James C Lester. 2018. Detecting and addressing frustration in a serious game for military training. *International Journal of Artificial Intelligence in Education* 28, 2 (2018), 152–193.

[25] Sidney D'Mello and Caitlin Mills. 2014. Emotions while writing about emotional and non-emotional topics. *Motivation and Emotion* 38 (2014), 140–156.

[26] Nigel Bosch and Sidney D'Mello. 2017. The affective experience of novice computer programmers. *International Journal of Artificial Intelligence in Education* 27 (2017), 181–206.

[27] Ma. Mercedes T. Rodrigo, Ryan S. Baker, Matthew C. Jadud, Anna Christine M. Amarra, Thomas Dy, Maria Beatriz V. Espejo-Lahoz, Sheryl Ann L. Lim, Sheila A. M. S. Pascua, Jessica O. Sugay, and Emily S. Tabanao. 2009. Affective and behavioral predictors of novice programmer achievement. In *Proc. of the 14th Annual ACM SIGCSE Conference*, 156–160.

[28] Kate Forbes-Riley, Mihai Rotaru, and Diane J. Litman. 2008. The relative impact of student affect on performance models in a spoken dialogue tutoring system. *User Modeling and User-Adapted Interaction* 18 (2008), 11–43.

[29] Maria Ofelia Z. San Pedro, Ryan S.J.d. Baker, Alex J. Bowers, and Neil T. Heffernan. 2013. Predicting college enrollment from student interaction with an intelligent tutoring system in middle school. In *Proc of the Int'l Conf on Educational Data Mining (EDM 2013).* 177–184.

[30] Nigel Bosch and Sidney D'Mello. 2014. Co-occurring affective states in automated computer programming education. In *Proc of the Workshop on AI-supported Education for Computer Science (AIEDCS) at the 12th Int'l Conf on Intelligent Tutoring Systems* (pp. 21-30).

[31] Harris Cooper and Larry V. Hedges. 1994. *Handbook of Research Synthesis, The.* Russell Sage Foundation.

[32] Maria Ofelia Z. San Pedro, Ryan S. Baker, and Neil T. Heffernan. 2017. An integrated look at middle school engagement and learning in digital environments as precursors to college attendance. *Technology, Knowledge and Learning* 22 (2017), 243–270.

[33] Michael Borenstein, Larry V. Hedges, Julian P.T. Higgins, and Hannah R. Rothstein. 2021. *Introduction to meta-analysis.* John Wiley & Sons.

[34] Cyril Bossard, Gilles Kermarrec, Cédric Buche, and Jacques Tisseau. 2008. Transfer of learning in virtual environments: a new challenge?. *Virtual Reality* 12 (2008), 151–161.

[35] Jon-Chao Hong, Ming-Yueh Hwang, Kai-Hsin Tai, and Pei-Hsin Lin. 2019. The effects of intrinsic cognitive load and gameplay interest on flow experience reflecting performance progress in a Chinese remote association game. *Computer Assisted Language Learning* 34, 3 (2019), 1–21.

[36] Cyril Brom, Michaela Buchtová, Vít Šisler, Filip Děchtěrenko, Rupert Palme, and Lisa Maria Glenk. 2014. Flow, social interaction anxiety and salivary cortisol responses in serious games: a quasi-experimental study. *Computers & Education* 79 (2014), 69–100.

[37] Victor Kostyuk, Ma. Victoria Almeda, and Ryan S. Baker. 2018. Correlating affect and behavior in reasoning mind with state test achievement. In *Proceedings of the 8th International Conference on Learning Analytics and Knowledge*. 26–30.

[38] Séverine Erhel and Eric Jamet. 2019. Improving instructions in educational computer games: Exploring the relations between goal specificity, flow experience and learning outcomes. *Computers in Human Behavior* 91 (2019), 106–114.

[39] Jon-Chao Hong, Kai-Hsin Tai, Ming-Yueh Hwang, and Yen-Chun Kuo. 2016. Internet cognitive failure affects learning progress as mediated by cognitive anxiety and flow while playing a Chinese antonym synonym game with interacting verbal–analytical and motor-control. *Computers & Education* 100 (2016), 32–44.

[40] Cyril Brom, Tereza Stárková, Edita Bromová, and Filip Děchtěrenko. 2019. Gamifying a simulation: Do a game goal, choice, points, and praise enhance learning?, *Journal of Educational Computing Research* 57, 6 (2019), 1575–1613.

[41] Hsin-Hsien Lu. 2016. Effects of competitive gaming scenario and personalized strategy on English vocabulary learning performance. Master's thesis, National Sun Yat-sen University.

[42] Arthur Graesser, Patrick Chipman, Brandon King, Bethany McDaniel, and Sidney D'Mello. 2007. Emotions and learning with AutoTutor. In *Proceedings of the 13th International Conference on Artificial Intelligence in Education*. 569–571.

[43] Maria Ofelia Z. San Pedro, Jaclyn L. Ocumpaugh, Ryan S. Baker, and Neil T. Heffernan. 2014. Predicting STEM and non-STEM college major enrollment from middle School interaction with mathematics educational software. In *Proceedings of the 7th International Conference on Educational Data Mining (EDM 2014)*. 276–279.

[44] T. F. Guia, M. M. Rodrigo, M. M. Dagami, J. O. Sugay, and F. J. Macam. 2013. An exploratory study of factors indicative of affective states of students using SQL-Tutor. *Research & Practice in Technology Enhanced Learning* 8, 3 (2013), 411–430.

[45] Cyril Brom, Vít Šisler, Michaela Slussareff, Tereza Selmbacherová, and Zdeněk Hlávka. 2016. You like it, you learn it: affectivity and learning in competitive social role play gaming. *International Journal of Computer-Supported Collaborative Learning* 11 (2016), 313–348.

[46] Chi-Cheng Chang, Clyde A Warden, Chaoyun Liang, and Guan-You Lin. 2018. Effects of digital game-based learning on achievement, flow and overall cognitive load. *Australasian Journal of Educational Technology* 34, 4 (2018), 155–167.

[47] Lilin Gong, Yang Liu, and Wei Zhao. 2019. Dynamics of emotional states and their relationship with learning outcomes during learning Python with MOOC. In *Proceedings of the 7th International Conference on Information and Education Technology*. 71–76.

[48] M. O. C. Z. San Pedro. 2016. Middle school learning, academic emotions and engagement as precursors to college attendance. Ph.D. Dissertation, Columbia University.

[49] Cyril Brom, Filip Děchtěrenko, Nikola Frollová, Tereza Stárková, Edita Bromová, and Sidney K. D'Mello. 2017. Enjoyment or involvement? Affective-motivational mediation during learning from a complex computerized simulation. *Computers & Education* 114 (2017), 236–254.

[50] Valerie J. Shute. 2011. Stealth assessment in computer-based games to support learning. In Computer Games and Instruction. Information Age Publishers, Charlotte, NC, 503-524.

[51] Sandra Becker. 2016. Situating emotions in the context of mathematics. Ph.D. Dissertation, University of Munich.

[52] Claudia Schrader and Slava Kalyuga. 2020. Linking students' emotions to engagement and writing performance when learning Japanese letters with a pen-based tablet: An investigation based on individual pen pressure parameters. *International Journal of Human-Computer*, 135 (2020), 102374.

[53] Jason M. Harley, Yang Liu, Tony Byunghoon Ahn, Susanne P. Lajoie, Andre P. Grace, Chayse Haldane, Andrea Whittaker, and Brea McLaughlin. 2019. I've got this: Fostering topic and technology-related emotional engagement and queer history knowledge with a mobile app. *Contemporary Educational Psychology* 59 (2019), 101790.

[54] Kate Forbes-Riley and Diane Litman. 2013. When does disengagement correlate with performance in spoken dialog computer tutoring?. *International Journal of Artificial Intelligence in Education* 22, 1–2 (2013), 39–58.

[55] Laura M. Naismith and Susanne P. Lajoie. 2018. Motivation and emotion predict medical students' attention to computer-based feedback. *Advances in Health Sciences Education* 23 (2018), 465–485.

[56] Shinobu Kitayama, Hazel Rose Markus, and Masaru Kurokawa. 2000. Culture, emotion, and well-being: Good feelings in Japan and the United States. *Cognition and Emotion* 14, 1 (2000), 93–124.

[57] John T. E. Richardson. 2004. Methodological issues in questionnaire-based research on student learning in higher education. *Educational Psychology Review* 16, 4 (2004), 347–358.

[58] Eda Okur, Sinem Aslan, Nese Alyuz, Asli Arslan Esme, and Ryan S. Baker. 2018. Role of socio-cultural differences in labeling students' affective states. in *Proceedings of the 19th International Conference on Artificial Intelligence in Education (AIED 2018)*. 367–380.

[59] Genaro Rebolledo-Mendez, N. Sofia Huerta-Pacheco, Ryan S. Baker, and Benedict du Boulay. 2021. Meta-affective behaviour within an intelligent tutoring system for mathematics. *International Journal of Artificial Intelligence in Education*.

[60] Nigel Bosch, Sidney D'Mello, Ryan Baker, Jaclyn Ocumpaugh, Valerie Shute, Matthew Ventura, Lubin Wang, and Weinan Zhao. 2015. Automatic Detection of Learning-Centered Affective States in the Wild. In *Proceedings of the 20th International Conference on Intelligent User Interfaces (IUI '15)*. 379–388.

[61] James J. Gross, Laura L. Carstensen, Monisha Pasupathi, Jeanne Tsai, Carina Götestam Skorpen, and Angie Y. C. Hsu. 1997. Emotion and aging: Experience, expression, and control. *Psychology and Aging* 12, 4 (1997), 590–599.

[62] Gregory A. Kimble. 1987. The scientific value of undergraduate research participation. *American Psychologist* 42, 3 (1987), 267–268.