A Re-Analysis and Synthesis of Data on Affect Dynamics in Learning

Shamya Karumbaiah, Ryan S. Baker, Jaclyn Ocumpaugh and Juliana Ma. Alexandra L. Andres

Abstract— Affect dynamics, the study of how affect develops and manifests over time, has become a popular area of research in affective computing for learning. In this paper, we first provide a detailed analysis of prior affect dynamics studies, elaborating both their findings and the contextual and methodological differences between these studies. We then address methodological concerns that have not been previously addressed in the literature, discussing how various edge cases should be treated. Next, we present mathematical evidence that several past studies applied the transition metric (L) incorrectly - leading to invalid conclusions of statistical significance - and provide a corrected method. Using this corrected analysis method, we reanalyze ten past affect datasets collected in diverse contexts and synthesize the results, determining that the findings do not match the most popular theoretical model of affect dynamics. Instead, our results highlight the need to focus on cultural factors in future affect dynamics research.

Index Terms— Education, Emotion in human-computer interaction, Emotion theory, Modeling human emotion

1 Introduction

Student affect in intelligent tutors 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 [1], analytical reasoning [2], motivation [3], and learning [4, 5]. Several research studies in the past decade have focused on building good quality affect detectors using physical and physiological sensors [6, 7, 8, 9], and interaction log data [10, 11, 12, 13]. Affect-sensitive interventions have been designed to improve student engagement [14], learning gains [5, 15, 16], and overall experience [17]. Developing effective real-time interventions depend on understanding how affect develops and manifests over time, an area of research termed *affect dynamics* (i.e. [18]), with a large body of research examining how students transition from one affective state to the next during learning activities (i.e., 2, 3, 5, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30].

The most commonly-cited model of affect dynamics in this context, put forward by D'Mello and Graesser [5], postulates that a specific set of affect transitions will be particularly prominent, but few empirical studies have matched that model's predictions, an issue which this paper investigates. 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 [31], appraise learning [32], and guide cognitive focus [22, 31, 33, 34]. These roles play a key function within the model [5] of affect dynamics during learning, which hypothesizes transitions between the educationally-important affective states of engaged concentration, confusion, frustration, and boredom (e.g., Fig. 1).

• *S. Karumbaiah, R.S. Baker, J. Ocumpaugh and J.M.A.L. Andres are with the Penn Center for Learning Analytics, University of Pennsylvania, 3700 Walnut St, Philadelphia, PA 19104, USA.*
E-mail: shamya@upenn.edu, ryanshaunbaker@gmail.com, jlocumpaugh@gmail.com, alexandralandres@gmail.com.

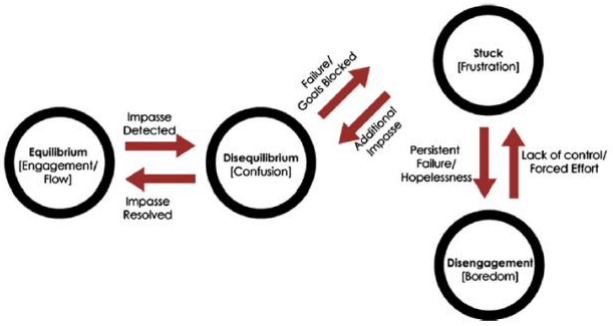


Fig. 1. Theoretical model of affect dynamics by D'Mello and Graesser [5]

The primary model cited from the paper predicts that students who experience an impasse during the flow state will transition to a state of disequilibrium, which manifests itself as the affective state of confusion. If the student resolves this impasse, they are predicted to transition back to flow. If, however, the impasse is not resolved, students are hypothesized to become "stuck" (experienced as frustration). If the frustration persists, the model suggests the learner will disengage, transitioning to boredom. Two other links in this highly cited model (*frustration → confusion* and *boredom→frustration*) are also hypothesized as likely, but the justification for these transitions is not discussed as thoroughly. This same paper also presented two empirical lab studies to demonstrate support for the hypothesized transitions. While the results aligned with the majority of the transitions in the model, neither of the studies supported the transition of *frustration→confusion* and one of them failed to support *frustration→boredom*.

D'Mello and Graesser's model has been widely referenced (with around 400 citations as of this writing) by various research studies on affect dynamics, including

many which have used the L likelihood statistic advanced in [5, 23]. Described in detail in section 3, the L statistic compares a specific transition's frequency to the frequency that might be expected based on its originating and destination affective states and can be used in statistical significance testing to infer whether a transition is more likely than chance. However, empirical results across a range of learning environments have not aligned with the theoretical model's proposed affective transitions (see Section 2).

A number of factors may be contributing to the divergence between the theoretical model and these empirical results. These studies have varied in terms of population (from elementary school students to graduate students and from the US to the Philippines), the learning context (type of learning activity as well as laboratory versus classroom study), and both study methodology and analysis method. However, another key difference between D'Mello and Graesser [5] and other research is how the data are represented when a student remains in the same affective state across several observation points. In [5], only transitions between differing states were considered, whereas in many other studies (including earlier work by the same authors), a student remaining in the same affective state was considered to exhibit a self-transition that was included in calculations. More broadly, past affect dynamics studies have differed in exactly how they calculate affect transitions, particularly in how they treated the edge cases (see Section 3.2).

In this paper, we first provide a detailed analysis of the prior affect dynamics studies elaborating on the contextual and methodological differences in them. We then describe the steps involved in affect dynamics analysis using L with clarifications on the edge cases that have mostly been omitted from write-ups on how the prior affect dynamics studies were conducted. Next, we present mathematical evidence that several past studies used the L statistic in ways that led to invalid conclusions of statistical significance and provided a correction to the interpretation of L statistic. Using a corrected analysis method, we re-analyze ten past affect datasets collected in diverse contexts and synthesize the results to find if there is empirical evidence for the D'Mello and Graesser's widely accepted model.

2 Prior Work on Affect Dynamics

To date, fifteen studies have used the L metric [23] to study the affect dynamics. The current study will focus primarily on the affective states included in the D'Mello and Graesser model (i.e., *boredom, engaged concentration, frustration, and confusion*), but as Table 1 summarizes, a range of other emotions have been included in these previously published papers (i.e., *anger, anxious, curiosity, delight, disgust, eureka, excitement, fear, happiness, neutral, sadness, and surprise*).

We use the term *engaged concentration* to refer to the affective state associated with flow [35], in line with the recommendations in [36], who noted that *flow* is a complex construct that goes beyond simply affective experience, also necessitating elements such as a perfect balance between challenge and ability. The reader should note that this state is alternately referred to in the affective dynamics literature as *flow*, *engagement, engaged concentration,* and *concentrating* due to different theoretical positions taken by the authors; however, the definitions used for this affective state are generally highly similar across papers.

TABLE 1
Affect States Studied in Previous Research on Affect Dynamics

Studies	BOR	ENG	DEL	FRU	SUR	NEU	CON	ANX	ANG	DIS	SAD	EUR	HAP	CUR	FEA	EXC
Andres & Rodrigo, 2014	x	x	x	x	x		x									
Botelho et al., 2018	x	x		x		x	x									
Baker, Rodrigo, & Xolocotzin, 2007	x	x	x	x	x	x	x									
Bosch & D'Mello, 2013	x	x		x			x									
Bosch, & D'Mello, 2017	x	x		x	x	x	x	x	x	x	x		x	x	x	
D'Mello & Graesser, 2012	x	x	x	x	x	x	x									
D'Mello et al., 2009	x			x	x	x	x	x	x	x	x	x	x	x		
D'Mello, Taylor, & Graesser, 2007	x	x	x	x	x		x									
D'Mello & Graesser, 2010	x	x	x	x	x		x									
Guia et al., 2011	x	x	x	x	x	x	x									
Guia et al., 2013	x	x	x	x	x	x	x									
McQuiggan et al., 2008; 2010	x	x	x	x			x	x	x		x				x	x
Ocumpaugh et al., 2017	x	x		x	x		x	x								
Rodrigo et al., 2008	x	x	x	x	x	x	x									
Rodrigo et al., 2011; 2012	x	x	x	x	x	x	x									

Categories studied in D'Mello & Graesser's Model are Highlighted in Gray. (BORed, ENGaged Concentration, DELight, FRUstration, SURprise, NEUtral, CONfused, ANXious, ANGer, DISgust, SADness, EUReka, CURious, FEAr, EXCited)

These studies have yielded a range of results. From the 15 studies considered, transitions that are both significantly more likely to occur than chance and align with the model of affect dynamics have been found predominantly in studies by D'Mello and his colleagues. *Engaged concentration→confusion* was reported in eight studies, including five by D'Mello et al. [4, 5, 21, 23, 37] as well as in studies by McQuiggan and colleagues [1, 26] and

Ocumpaugh and colleagues [28]. However, fewer studies found support for other predicted transitions. *Confusion→engaged concentration* was reported in three D'Mello et al. studies [4, 5, 37] and in one study by Ocumpaugh et al. [28] and Botelho et al. [38]. Transitions of *confusion→frustration* (in [4, 5, 21, 37]) and *frustration→confusion* (in [4, 2, 21]) were reported exclusively in studies by D'Mello and his colleagues. *Frustration→boredom* was reported in D'Mello et al. studies [4, 5] and was marginally significant in one Rodrigo et al. study [3]. *Boredom→frustration* was reported in two studies by D'Mello et al. [5, 23] and in one study by Rodrigo and colleagues [30] and Botelho et al. [38].

2.1 Demographic Differences in Previous Work Examined

Across all of the hypothesized affective transitions, only one transition is seen in more than half of the studies, arguing that thus far, the theoretical predictions of this model are not being upheld. However, the 15 studies summarized in Table 2 differ noticeably in terms of the demographic characteristics of their samples, including age and the region where the research was conducted. It is possible that these differences may explain the inconsistencies in whether research supports the model.

TABLE 2
SUMMARY OF THE OBSERVED METHODOLOGICAL DIFFERENCES ACROSS 15 STUDIES ON AFFECT DYNAMICS

	Region	Age	N	School/Grade Population	Learning System	Class v. Lab	Obs. Type/ Grain Size	Obs. Session	Self-trans	Aligned Transitions
Andres & Rodrigo, 2014	Quezon City, PH	13-16	60	Public school	Physics Playground	C	QFO	2hrs	Inc	0
Botelho et al., 2018	--	--	838	--	ASSISTments	C	Automated Detector. 20s	--	Exc	--
Baker et al., 2007	Manila, PH	14-19	36	High school	Inc. Machine	C	QFO ev. 60s	10min	Inc	0
Bosch & D'Mello, 2013	US	--	29	Undergrads	Unnamed	L	RJP on 100 fixed points	25min	Exc	3
Bosch, & D'Mello, 2017	Midwestern US	17-21	99	Undergrads	Unnamed	L	RJP on 100 fixed points	25min	Exc	5
D'Mello & Graesser, 2012	Southern US	--	28; 30	Undergrads	Auto-Tutor	L	RJP every 20s; fixed points	32min; 35min	Exc	4;5
D'Mello et al., 2007	Southern US	--	28	Undergrads	Auto-Tutor	L	RJP ev. 20s	32min	Inc	2
D'Mello et al., 2009	Southern US	--	41	Undergrads	Unnamed	L	RJP on fixed points	35min	Exc	1
D'Mello & Graesser, 2010	Southern US	--	28; 30	Undergrads	Auto-Tutor	L	RJP ev. 20s; fixed points	32min; 35min	Exc	3;3
Guia et al., 2011; 2013	Quezon City, PH	18-20	60	Undergrads	SQL Tutor	C	QFO ev. 200s	1hr	Inc	0
McQuiggan et al., 2008; 2010	US	21-60	35	Grad students	Crystal Island	L	SRI	35min	Inc	1
Ocumpaugh et al., 2017	New York, US	18-22	108	West Point	vMedic	C	QFO ev.122s	--	Inc	2
Rodrigo et al., 2008	Quezon City & Cavite Prov., PH	9-13	180	Private school	Ecolab	C	QFO	40min	Inc	1
Rodrigo et al., 2011; 2012	Quezon City, PH	43813	126	High school	Scatterplot Tutor	C	QFO ev. 200s	80min	Inc	1

** PH: Philippines, QFO: Qualitative field observation, RJP: Retrospective judgment protocol, SRI: self-report based on interactions, Inc: self transitions included, Exc: self transitions excluded.*

All the studies by D'Mello and colleagues were conducted in the United States with undergraduate populations. By contrast, the studies by other researchers come from a wider range of demographics with students in middle school (private), high school (public and private), undergraduate programs, and graduate schools, from locations in the United States and in the Philippines. Differences in culture are known to influence variation in beliefs and personal dispositions towards emotional expression and moderation [39] and the frequency and emergence of certain affective states [40]. Likewise, age is known to influence emotional expressivity [41, 42] and inhibition [43]. It is possible that some of the differences in results may be due to these factors; if so, this would suggest that D'Mello and Graesser's model may not be generalized across contexts.

2.2 Learning Settings

The studies were conducted across multiple instructional settings, including regular classroom environments and laboratory settings. Educational software used to investigate affective dynamics has covered a broad range of educational content, including mathematics [29, 30], biology [3, 26, 27], emergency medical practices [28], physics [19, 23, 44], computer literacy, and programming [4, 21, 24, 25, 37], and analytical problem solving [2]. The learning systems that have been used across these studies have also differed in terms of design. The Scatterplot Tutor, SQL-Tutor, AutoTutor and the other researcher-built learning environments used in studies conducted by D'Mello follow more linear designs wherein learners must complete problems before they are able to proceed. On the other hand, environments such as Physics Playground, Crystal Island, The Incredible Machine, vMedic, and Ecolab, are more open-ended systems that offer learners the opportunity to explore the range of possible solutions.

2.3 Data Collection Procedure, including Observation Grain-Size and Session Duration

Six of the 15 studies use the Baker Rodrigo Ocumpaugh Monitoring Protocol (BROMP, [45]), a momentary time sampling method that uses a holistic coding practice to code for both affect and behavior. In this protocol, students are observed for up to 20 seconds in a round-robin manner throughout the given observation period to ensure uniform frequencies of student observation. The protocol is enforced by an Android application known as the Human Affect Recording Tool (HART, [45]).

By contrast, D'Mello and his colleagues have used self-reporting methods, collecting affect data through retrospective judgment protocols which synchronize webcam video of students' faces to screen capture of the learning environment [2, 4, 5, 21, 23, 37]. McQuiggan and his colleagues also collected self-reported data, but used in-game dialogs to collect spontaneous reports rather than using a retrospective technique [26, 27].

Observation sessions in this research varied in length, ranging from 10 minutes [20] to 2 hours [19], potentially influencing the affect that emerges during observation. Prolonged exposure to similar tasks may produce fatigue or boredom [46], decreasing learner performance [47]. Correspondingly, it may also increase students' susceptibility to what [23] describes as vicious cycles of boredom, where learners are unable to transition to other affective states.

2.4 Differences in the Treatment of Self-Transitions Between Studies

All of the studies considered in this section analyze time-series data (e.g., the order of the occurrences of each affective state), but they have been inconsistent in their treatment of "self-transitions," which occur when a student remains in the same affective state over two (or more) consecutive observations. In more recent studies, D'Mello and colleagues have removed self-transitions during the data preparation stage [2, 4, 5, 21, 37], as have Botelho and colleagues [38]. For example, a sequence of *confusion, frustration, frustration, boredom* has one self-transition (from *frustration* to *frustration*), and would be modified into *confusion, frustration, boredom*.

However, this practice is not followed in all work. Nearly a dozen other studies conducted in this field do not report discarding self-transitions in their data processing [3, 24, 25, 26, 27, 28, 29, 44], including work by D'Mello and his colleagues (e.g. [23]). The choice of including or excluding self-transitions is likely based on the goal of the research study; including self- transitions may suppress non-self transitions. If some affective states are particularly persistent [36], including self-transitions in analysis helps better understand each state's persistence, but dilutes evidence for transitions between different affective states. In contrast, excluding self-transitions could be a better choice if the goal is to reveal a larger number of affective patterns that might otherwise be suppressed by the presence of persistent affective states. However, as we discuss in sections 3.3 and 3.4, this seemingly small step may have disproportionate effects on study outcomes, particularly in terms of this decision's impact on the interpretation of commonly used statistics.

2.5 Summary of Past Results

Tables 3 and 4 present the findings of the affect dynamics studies conducted in the past. Note that blank cells in Table 3 represent transitions that were either not studied or not reported in each paper. These transitions are also not counted in the calculations in Table 4. Also, Table 3 does not include [29] as this study only reported self-transitions. The studies that did not report discarding self-transitions predominantly reported null effects for the non-self transitions. (See Table 3, column 6.)

The proportion of studies that report significant, positive likelihood for non-self transitions, and in particular for the transitions hypothesized by the D'Mello and Graesser model, is higher in the set where the self-transitions were discarded. This could imply that discarding self-transitions can result in higher conformance with the D'Mello and Graesser model. But, currently, all the studies (except one) in the set where self-transitions are discarded correspond to studies conducted by D'Mello and colleagues in lab settings involving undergraduate population from the United States and use retrospective affect judgments to collect affect data in observation sessions that are around 30 minutes long.

Previous work [48] removed self-transitions and reanalyzed data collected in a classroom setting in the Philippines, where 180 eighth and tenth graders used a learning game called Physics Playground [49] for 2 hours. In this study, affect data was collected through field observations using the BROMP protocol [45]. This paper found that excluding self-transitions increased the proportion of transitions that occur above chance, yet it did not lead to having a larger number of

transitions that were more likely than chance and conformed to D'Mello and Graesser's theoretical model.

TABLE 3
SIGNIFICANCE OF THE TRANSITIONS REPORTED IN PREVIOUS RESEARCH

Studies	ENG_ENG	ENG_CON	ENG_FRU	ENG_BOR	CON_ENG	CON_CON	CON_FRU	CON_BOR	FRU_ENG	FRU_CON	FRU_FRU	FRU_BOR	BOR_ENG	BOR_CON	BOR_FRU	BOR_BOR
Andres & Rodrigo, 2014	+	Ø	Ø	-	Ø	Ø	Ø	Ø	-	Ø	+	Ø	-	-	Ø	+
Baker, Rodrigo, & Xolocotzin, 2007	+	-	Ø	-	Ø	+	Ø	Ø	Ø	Ø	Ø	Ø	Ø	Ø	Ø	+
Botelho et al., 2018		Ø	-	+	+		-	-	+	Ø		-	+	Ø	+	
Bosch & D'Mello, 2013		+			Ø		+			+		Ø	-		Ø	
Bosch, & D'Mello, 2017		+	Ø	Ø	+		+	Ø	Ø	+		+	+	Ø	Ø	
D'Mello & Graesser, 2012 [Study 1]		+	Ø	Ø	+		+	Ø	Ø	Ø		Ø	Ø	Ø	+	
D'Mello & Graesser, 2012 [Study 2]		+	Ø	Ø	+		+	Ø	+	Ø		+	Ø	Ø	+	
D'Mello et al., 2009										+						
D'Mello, Taylor, & Graesser, 2007	+	+	-	-	Ø	+	Ø	-	Ø	Ø	Ø	Ø	Ø	Ø	+	+
D'Mello & Graesser, 2010 [Study 1]		+			+		+									
D'Mello & Graesser, 2010 [Study 2]		+			+		+									
Guia et al., 2011	Ø	Ø	Ø	Ø	-	Ø	-	Ø					Ø	Ø	-	Ø
Guia et al., 2013	Ø	Ø	Ø	Ø	-	Ø	Ø	Ø	-	Ø	+	Ø	Ø	-	-	+
McQuiggan et al., 2008; 2010	Ø	+	-	Ø	Ø	Ø	Ø	Ø	Ø	Ø	Ø	Ø	Ø	Ø	Ø	Ø
Ocumpaugh et al., 2017	-	+	Ø	+	+	-	Ø	Ø	Ø	Ø	-	-	Ø	+	Ø	-
Rodrigo et al., 2008 [Control]	+	Ø	Ø	-	Ø	Ø	Ø	Ø	-	Ø	+	+	-	Ø	Ø	+
Rodrigo et al., 2008 [Experiment]	Ø	Ø	Ø	-	Ø	Ø	Ø	Ø	-	Ø	+	Ø	-	-	Ø	+
Rodrigo et al., 2011 [Control]	+					+					Ø					+
Rodrigo et al., 2011 [Experiment]	+					+					Ø					+
Rodrigo et al., 2012 [Control]	+	-	Ø	Ø	-	+	Ø	-	Ø	Ø	Ø	Ø	Ø	-	+	+
Rodrigo et al., 2012 [Experiment]	+	-	+	-	-	+	Ø	-	Ø	Ø	Ø	Ø	Ø	-	+	+

+ indicates a significant positive transition, - indicates a significant negative transition and Ø indicates an insignificant transition. Transitions not studied or not reported are left blank. A transition from affect1 to affect2 is denoted as "affect1_affect2." For instance, CON_BOR is a transition from confusion to boredom. Only the states from D'Mello model are included - ENGaged concentration, CONfused, FRUstration, BORed. D'Mello studies and transitions studied in D'Mello & Graesser's model are highlighted in grey.

TABLE 4
PROPORTION OF STUDIES THAT REPORTED POSITIVE, NEGATIVE, AND NULL VALUES FOR THE NON-SELF TRANSITIONS

Transition	The proportion of studies reporting a significant **positive** L value		The proportion of studies reporting a significant **negative** L value		The proportion of studies reporting a **non-significant** L value	
	INC	EXC	INC	EXC	INC	EXC
ENG_CON	0.25	1.00	0.25	0.00	0.50	0.00
ENG_FRU	0.08	0.00	0.25	0.00	0.67	1.00
ENG_BOR	0.17	0.00	0.50	0.00	0.33	1.00
CON_ENG	0.17	0.83	0.33	0.00	0.50	0.17
CON_FRU	0.00	1.00	0.17	0.00	0.83	0.00
CON_BOR	0.00	0.00	0.33	0.00	0.67	1.00
FRU_ENG	0.09	0.33	0.36	0.00	0.54	0.67
FRU_CON	0.00	0.60	0.00	0.00	1.00	0.40
FRU_BOR	0.09	0.50	0.18	0.00	0.72	0.50
BOR_ENG	0.08	0.25	0.25	0.25	0.67	0.50
BOR_CON	0.08	0.00	0.42	0.00	0.50	1.00
BOR_FRU	0.33	0.50	0.17	0.00	0.50	0.50

INC - Studies that includes self-transitions total = 12); EXC - Studies that exclude self-transitions (total = 7). The transitions that were either not studied or not reported are not counted for that study. Transitions studied in D'Mello & Graesser's model are highlighted in grey.

3 THE TRANSITION METRIC *L*

In this section, we present the metric most commonly used in affect dynamics research, the ***L*** statistic [5], and provide our recommendations for how to handle the several special cases that occur in using this metric. Next, we discuss issues surrounding the underlying assumptions of this metric, provide mathematical evidence that the metric has been used in an invalid fashion in many past papers, and propose a correction for future use.

3.1 The L Statistic for Affect Dynamics

Given an affect sequence, the ***L*** statistic [23] calculates the likelihood that an affective state (*prev*) will transition to a subsequent (*next*) state, given the base rate of the next state occurring.

$$L(prev \rightarrow next) = \frac{P(next|prev) - P(next)}{1 - P(next)} \qquad (1)$$

The expected probability for an affective state, *P(next),* is the percentage of times that the state occurred as a next state. Thus, the first affective state in a student's sequence will be excluded from this calculation since this state cannot take the

role of a next state. For instance, for a state sequence AABB, the probability of the state A as the next state, P(A_next) is 0.33 (from ABB) instead of 0.5. Similarly, the calculation of the *prev* state excludes the last state in the sequence. The conditional probability, *P(next|prev)* is given by:

$$P(next \mid prev) = \frac{Count(prev \rightarrow next)}{Count(prev)} \quad (2)$$

where *Count(prev → next)* is the number of times the *prev* state transitioned to the next state, and *Count(prev)* is the number of times the state in *prev* occurred as the previous state. In the example sequence of AABB, Count(B_prev) is 1 (from AAB) instead of 2 as the last state in the sequence cannot be a *prev* state for any transition.

The value of $\mathcal{L}$ varies from -∞ to 1. D'Mello and Graesser [5] state that "the sign and the magnitude of $\mathcal{L}$ is intuitively understandable as the direction and size of the association." As has been expanded in subsequent papers [2, 4, 5, 19, 21, 24, 25, 26, 27, 29, 30, 37, 48, 50, 51], $\mathcal{L}$ = 0 is treated as chance, while $\mathcal{L}$ > 0 and $\mathcal{L}$ < 0 are treated as transitions that are more likely or less likely (respectively) than chance.

To perform affect dynamics analysis across all students in an experiment, first the $\mathcal{L}$ value for each affect combination is calculated individually per student. Next, as [5, pg. 7] recommends, the researcher runs "one-sample [two-tailed] t-tests to test whether likelihoods were significantly greater than or equivalent to zero (no relationship between immediate and next state)," on the sample of individual student $\mathcal{L}$ values for each transition.

Lastly, a Benjamini-Hochberg post-hoc correction procedure is used by some of the research groups conducting this type of analysis [3, 19, 29, 30, 38, 48, 51] to control for false-positive results since the set of hypotheses involves multiple comparisons – however, some early research papers by these groups omitted this step, and other research groups have not used any type of post-hoc correction at all.

3.2 Special Cases when Implementing L

There are several special cases in the calculation of $\mathcal{L}$ where there is no consensus in the literature on how to perform the calculation. [48] has recommended the following treatment:

1. When any affective state (A_n) being considered in a given study is not present for a given student's observation period:
   a. If transitions to A_n do not occur for that student, then *P(next)* = 0 and *P(next | prev)* = 0, and thus, $\mathcal{L}$ = 0.
   b. If transitions from A_n also do not occur, then we do not know what affective state would have followed A_n, and thus, $\mathcal{L}$ = undefined.
2. Following from case 1, if a student remains in a single affective state (A_s) throughout an observation period, the calculations differ based on whether or not the self-transitions are included.
   a. If self-transitions are included in the analyses, then:
      i) Transitions from A_s to any other affective state (e.g., A_n) do not occur, and therefore, as in 1a, $\mathcal{L}$ = 0 for any transition out of A_s.
      ii) Transitions to A_s from any other affective state (e.g., A_n) do not occur, and therefore, as in 1b, $\mathcal{L}$ = undefined.
   b. If self-transitions are discarded in the analyses, an affect sequence consisting of repeated observations of the same affective category is reduced to a single observation of that affective state. In this case, no transitions occur, and therefore $\mathcal{L}$ = undefined for all possible sequences being studied.

It is not always clear how these special cases are treated in the past published research. In this study, we follow [48]'s definition of $\mathcal{L}$ as outlined above.

3.3 The Case of Self-Transitions

One other special case that is not fully discussed in most of the literature is the case of self-transitions. In the majority of articles written by D'Mello's group and other groups' articles as well [i.e., 38], self-transitions (where the student remains in the same affective state for more than one step in a sequence) are removed. This straightforward procedure seems quite logical, but there is evidence that something may be wrong with the resulting calculations. Notably, in [38], after removing self-transitions, all transitions into the affective state of engaged concentration were more likely than chance. As such, it may be worth examining the mathematical assumptions of this procedure. Specifically, while calculating the transition likelihood from the affective state of M_t (*prev*) to M_{t+1} (*next*), D'Mello explains that, "...if M_{t+1} and M_t are *independent* [emphasis added], then $Pr(M_{t+1}|M_t) = Pr(M_{t+1})$" [15]. However, removing self-transitions violates the assumption of independence between M_{t+1} and M_t, as M_{t+1} can now only take values other than M_t. For instance, if there are three states (A, B, C) and if M_t = A, then M_{t+1} can only take the value of either B or C if self-transitions are not allowed. Hence, when self-transitions are excluded, $Pr(M_{t+1}|M_t) \neq Pr(M_{t+1})$ when M_t and M_{t+1} are independent.

Another sign of potential problems is found in [5], when that paper draws an analogy between $\mathcal{L}$ statistics and Cohen's kappa, saying, "The reader may note significant similarity to Cohen's kappa for agreement between raters and indeed the likelihood metric can be justified in a similar fashion." Although this analogy seems compelling based on the similarity of the equations between the two metrics, it is worth noting that there is a difference between the range of values the two statistics can take. While the value of $\mathcal{L}$ varies from -∞ to 1 [44], the value of Cohen's kappa varies from -1 to 1 [66].

These findings raise the question: if a transition occurs at chance, and self-transitions are excluded, is the value of $\mathcal{L}$ still 0? We address this concern in section 3.4.

3.4 Correcting Chance L Value

For a state space with n affective states (n > 2), there would be n^2 unique transitions if we include self-transitions, but only $n^2 - n$ unique transitions if we exclude self-transitions [50]. Thus, at chance, the expected probability is

$$P(next) = \frac{n}{n^2} = \frac{1}{n} \quad \text{if self-transitions are included}$$

$$P(next) = \frac{n-1}{n^2-n} = \frac{1}{n} \quad \text{if self-transitions are excluded}$$

However, at chance, the conditional probability is

$$P(next \mid prev) = \frac{1}{n} \quad \text{if self-transitions are included}$$

$$P(next \mid prev) = \frac{1}{n-1} \quad \text{if self-transitions are excluded}$$

Plugging these into the original equation of L (equation 1), the value of L at chance is

$$L = 0 \quad \text{if self-transitions are included}$$

$$L = \frac{1}{(n-1)^2} \quad \text{if self-transitions are excluded}$$

This finding shows that the L statistic must be interpreted differently depending on how many affective categories are being observed. Table 5 shows the values at chance, depending on how many affective states are being observed.

TABLE 5
THE VALUE OF L THAT REPRESENTS CHANCE, FOR VARYING STATE SPACE

n	3	4	5	6	7	8
chance L	0.25	0.11	0.0625	0.04	0.0277	0.0204

As noted above, affect dynamics is most frequently studied in terms of four or five affective states. In such a setup (n = 5), the L value at chance is L=0.0625. For the smallest reasonable state space (n = 3), the L value at chance reaches 0.25. As the number of affective states observed increases, the impact of the difference between including and excluding self-transitions decreases (Table 5). This is particularly a problem if a statistical significance test is conducted that compares to a chance value of 0. Take, for instance, a case where three affective states are studied, and L is reliably 0.15 for a specific transition. In this case, a comparison to 0 may find that a transition occurs more often than chance when it actually occurs less often than chance.

4 RE-ANALYSIS OF PRIOR DATA

4.1 Datasets used in this analysis

In order to collect affect datasets from diverse contexts, we reached out to authors of papers that have previously reported work on affect in learning and obtained 10 datasets. Figure 2 shows pictures of the eight online learning systems involved in several of these studies, and Table 6 provides an overview of these datasets, which are described in greater detail in the following subsections. Two of the datasets were collected in classroom studies with no learning system. In addition to providing information about the learning setting involved in each study, these sections (4.1.1-4.1.10) also outline the demographics of the participants in the study.

Nine of these datasets were produced using the BROMP protocol [44] to collect affect in classroom settings. BROMP is a momentary time sampling method where students are briefly observed by certified coders one after another, in repeated round-robin cycles. BROMP has been used by over 160 researchers and practitioners in seven countries for field observations.

There are slight variations in how BROMP may be implemented in the different countries that are represented in this study. BROMP observations in the Philippines have historically used two coders making simultaneous observations on the same student. In contrast, BROMP coders in the United States record observations independently (after inter-rater checks are complete). In the former case, the observations from the different coders are merged to form a single affect sequence sorted by the time of observation.

4.1.1 Dataset #1- ASSISTments

Affect data from 856 students was collected in six schools across urban, suburban, and rural settings [3]. A total of 7,663 field observations were collected. ASSISTments (Figure 2a) is a free web-based platform which is used by students in the classroom and at home to practice the learning content assigned by their teacher. It is designed to provide immediate feedback and on-demand hints and sequences of scaffolding to support students when they make errors.

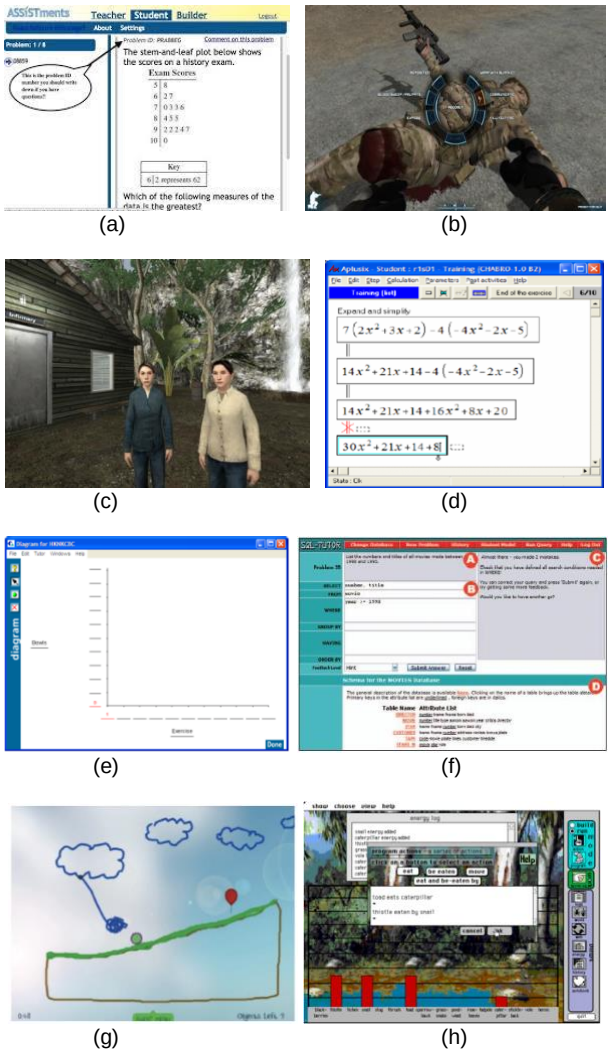


Fig. 2. Snapshots of the different learning systems studied in this work. In parenthesis is the dataset number. From top left - a) ASSISTments (#1); b) vMedic (#4); c) Crystal Island (#5); d) Aplusix (DS#6); e) Scatterplot (#7); f) SQL-Tutor (#8); g) Physics Playground (#9); h) Ecolab (#10).

TABLE 6
DESCRIPTION OF THE 12 DATASETS REANALYZED IN THIS PAPER

Dataset#	Learning Sys/ Classroom	No. of Students	No. of Coders at Once	No. of Observations Collected	No. of Students Dropped	No. of Observations				
						ENG	CON	FRU	BOR	NA
1	Assistments	856	1	7673	167	5039	469	239	657	1269
2	Classroom	371	1	3793	58	2957	27	27	618	191
3	Classroom	18	1	5308	0	4012	38	2	925	331
4	vMedic	117	1	755	32	435	174	32	73	41
5	Crystal Island	35	0	592	6	249	77	35	19	212
6	Aplusix	140	2	3640	7	2641	494	101	126	278
7A	Scooter Control	61	2	2976	0	1293	1406	13	179	85
7B	Scooter Experiment	64	2	3072	0	1142	1605	8	204	113
8	SQL Tutor	29	2	1044	0	481	388	39	53	83
9	Physics Playground	241	2	62502	0	46040	3962	3469	2557	6474
10 A	Ecolab Control	90	2	4560	2	2855	638	235	596	236
10 B	Ecolab Experiment	90	2	4560	0	3164	621	131	420	224

4.1.2 Dataset #2- Elementary Classes

Godwin and colleagues [53] conducted observations across twenty-two classrooms that were selected from 5 charter schools located in or near a medium-sized Northeastern city in the United States of America. Students observed were between kindergarten and fourth grade. The average class size was 21 students (10 males, 11 females). The number of children observed per session ranged from 15 to 22 children. The observation sessions were staggered across three time periods, and a total of 128 observation sessions were conducted in their study. Each observation session lasted approximately one hour. The average number of observations per session was 346.13, and the average number of observations per student within a session was 19.27.

4.1.3 Dataset #3- Graduate Level Classes

DiStefano and colleagues [54] observed students in an introductory methods course in an urban graduate school of education in the Northeast, United States. There are approximately 25 full-time and 408 part-time Graduate School of Education students (81% female and 19% male; 53% White, 19% Black, 11% Asian, 16% unknown, & 1% Hawaiian). Students participating in the course were observed and coded across four different classroom conditions of class lecture, class discussion, small group work, and transition periods for one semester.

4.1.4 Dataset #4 - vMedic

Ocumpaugh and her colleagues [28] collected affect data from 108 West Point cadets (ages of 18-22) using vMedic (a.k.a. TC3Sim). vMedic (Figure 2b) is a virtual world developed for the US Army to provide training in combat medicine and battlefield doctrine around medical first response. Two BROMP-certified coders observed the trainees while they used vMedic for up to 25 minutes. The coders observed different trainees at a given time, coding for surprise and anxiety as well as more commonly-studied affective states. Each trainee was observed once every 122 seconds (std dev = 100.14), leading to a total of 756 observations.

4.1.5 Dataset #5 - Crystal Island

Affect data in crystal island environment (Figure 2c) was collected by McQuiggan and colleagues [1, 26], where they observed 35 graduate students ranging in age from 21 to 60 (M = 24.4, SD = 6.41). This included 9 females and 26 males, and 60% were Asian (n = 21) and approximately 37% were Caucasian (n = 13). Participants interacted with crystal island for 35 minutes and self-reported their affective state via an in-game dialog from a selection of ten affective states (anger, anxiety, boredom, confusion, delight, excitement, fear, engaged concentration, frustration, and sadness). A total of 592 self-report of affective states was collected. Crystal Island is a game-based learning environment designed for middle-school students in the domains of microbiology and genetics to develop deeper understandings about scientific knowledge.

4.1.6 Dataset #6 - Aplusix

Rodrigo and colleagues [29] collected data from 140 high school students (ages 12-15; 83 females, 57 males) using Aplusix in 2008 within four schools within Metro Manila and one school in the Province of Cavite (Philippines). Students used Aplusix II [55], an algebra learning assistant that teaches students how to balance equations (Figure 2d). In each session, ten students were observed for 45 minutes. Each student has 13 observations spaced three minutes apart for a total of 3640 observations across all students.

4.1.7 Dataset #7a and 7b – Scatterplot Tutor

Scatterplot Tutor data from 125 students (ages 12-14) was collected in 2008 from an urban high school in Quezon City in the Philippines during a project with both a control (7A) and an experimental (7B) condition [29, 30]. Scatterplot Tutor (Figure 2 e) is a Cognitive Tutor that teaches the generation and interpretation of scatter plots. Both datasets 7A and 7B were collected from a group of ten students over 80-minute learning sessions yielding 24 observations per student. The control group consisted of 61 students, yielding a total of 1,464 observations, while the experimental group consisted of 64 students, resulting in a total of 1,536 observations. In dataset 7B, the experimental condition, an interactive pedagogical agent named Scooter designed to reduce gaming the system behavior, was shown on-screen while the students interacted with the learning system. For the purpose of this study, data from the control group and the experimental group are treated as two separate datasets (7A and 7B) in order to account for any influences Scooter may have had on the emergence of affective states.

4.1.8 Dataset #8 – SQL-Tutor

Guia and colleagues [24] collected data on affect from 29 third-year undergraduate students from Ateneo de Manila University in the Philippines while using SQL-Tutor. SQL-Tutor [27] (Figure 2f) is an intelligent tutor that is designed to teach

Structured Query Language (SQL). The participants of this study were in a course that required knowledge in database programming, but none had previously used SQL-Tutor. The participants were randomly divided into three sections and were asked to use SQL-Tutor for 60 minutes. Each student was observed once per 200 seconds leading to a total of 1044 observations.

4.1.9 Dataset #9 - Physics Playground

The Physics playground data was collected in 2015 from 180 students: 120 8th-graders and 60 10th-graders from Baguio, Cebu, and Davao, in the Philippines [49]. Students spent 2 hours using Physics Playground (Figure 2g), a learning environment that teaches qualitative physics to secondary students [56]. In this 2-dimensional game, students sketch different objects like pendulums, ramps, levers, and springboards to guide a ball to touch a balloon. Laws of physics apply to all the objects on the screen. Each student was observed approximately once per minute. On average, there were 135 observations per student, giving a total of 24,330 observations.

4.1.10 Dataset #10a and 10b - Ecolab

Rodrigo and colleagues [3] collected affect data from 180 students from two private, co-educational grade schools in the Philippines (ages 9-13) while they used the Ecolab learning system (Figure 2h) to learn about food webs and chains. There were ten students per observation session, five in control (Ecolab) and five in experimental (M-Ecolab) condition. In M-Ecolab, the system was enhanced with an affective learning companion who modified its demeanor based on automated assessments of the learner's degree of motivation. Students used the system for 40 minutes, and each student was observed for affect 12 times using BROMP.

In the current study, this dataset is split between the control (10A) and experiment conditions (10B) and used separately for the analysis. Each of the sub-datasets consists of 90 students and contains a total of 4560 observations across all students.

4.2 Affect Distribution Across Datasets

The descriptive statistics on the distribution of the affective states across these states are given in Table 7. In order to be consistent, affective states other than the four theorized in the D'Mello and Graesser model have been converted to a not-applicable (N/A) label. Overall, across datasets, there was a relatively high incidence of *engaged concentration* followed by *confusion* and *boredom,* and there was a relatively low incidence of *frustration*.

TABLE 7
MEAN AND STANDARD DEVIATION OF THE PROPORTIONS OF THE AFFECTIVE STATES ACROSS THE STUDENTS IN THE 12 DATASETS REANALYZED IN THIS PAPER

Dataset#	Learning Sys. / Classroom	ENG	CON	FRU	BOR	NA
1	Assistments	0.628 (0.289)	0.068 (0.13)	0.03 (0.078)	0.097 (0.171)	0.177 (0.217)
2	Classroom	0.781 (0.176)	0.007 (0.028)	0.007 (0.028)	0.162 (0.143)	0.05 (0.101)
3	Classroom	0.756 (0.056)	0.007 (0.005)	0 (0.002)	0.174 (0.061)	0.062 (0.017)
4	vMedic	0.572 (0.315)	0.247 (0.286)	0.047 (0.103)	0.09 (0.157)	0.043 (0.102)
5	Crystal Island	0.453 (0.334)	0.118 (0.124)	0.058 (0.091)	0.03 (0.056)	0.342 (0.287)
6	Aplusix	0.726 (0.185)	0.136 (0.107)	0.028 (0.057)	0.035 (0.076)	0.076 (0.087)
7A	Scooter Control	0.437 (0.252)	0.469 (0.221)	0.004 (0.015)	0.061 (0.15)	0.028 (0.043)
7B	Scooter Experiment	0.372 (0.207)	0.522 (0.181)	0.003 (0.01)	0.066 (0.145)	0.037 (0.058)
8	SQL-tutor	0.461 (0.166)	0.372 (0.168)	0.037 (0.072)	0.051 (0.101)	0.08 (0.085)
9	Physics Playground	0.74 (0.147)	0.062 (0.062)	0.059 (0.071)	0.032 (0.062)	0.107 (0.101)
10A	Ecolab Control	0.624 (0.21)	0.136 (0.11)	0.054 (0.094)	0.142 (0.178)	0.044 (0.079)
10B	Ecolab Experiment	0.675 (0.188)	0.135 (0.117)	0.031 (0.071)	0.115 (0.16)	0.043 (0.073)

Standard deviations presented in paretheses.

5 METHODS

This study reanalyzes the 12 datasets outlined in the previous section. Specifically, we standardize the treatment of transition types and edge cases that have been identified as sources of potential discrepancies in the results between studies. This allows us to compare the results for individual datasets to determine which show the most conformity to the D'Mello & Graesser model. Then we apply Stouffer's Z, a method that allows us to summarize results across multiple datasets.

5.1 Standardizing the Analysis of Transition Types and Edge Cases

As discussed in section 3, in studies that included self-transitions, the results for non-self transitions were less likely to be positive in direction and less likely to be statistically significant. While understanding the persistence of affective states might be important in practice (algorithms designed to trigger interventions, for instance), focusing on out-of-state transitions could be more important for a theoretical model of affect dynamics. As such, we have decided to exclude self-transitions in this work, and we have reanalyzed these datasets accordingly.

In addition, we have standardized our treatment of edge cases. Specifically, we have ensured that in cases where a student remains in the same affective state throughout the observation session, we discard the student from the analysis. The number of students who remained in a single affective state and were omitted from the analysis is given in Table 6.

5.2 Stouffer's Z to Summarize Significance Levels from Multiple Affect Datasets

In order to determine whether transitions are significantly more likely than chance across datasets, we combine p-values from the independent significance tests conducted on the multiple affect datasets, using Stouffer's Z [57], also known as the sum of Z's method, a classic method for summarizing significance values in the social sciences [58] where the datasets do not contain any of the same subjects (which we believe to be true of these datasets). For the k independent tests (k = number of affect datasets), Stouffer's Z is given by

$$\sum_{i=1}^{k} z(p_i)/\sqrt{k} \quad (1)$$

where, p_i is the p-value from the i[th] affect dataset. This statistic can then be used in a Z statistical test. This is repeated for all the 12 non-self-transitions being studied in this paper. Since the *L* values can take both positive and negative values, we are using the two-tailed version of Stouffer's Z. For negative *L* values, the corresponding Z scores are converted to a negative value. This method looks across all tests to see what the aggregate evidence is in favor of there being a significant relationship. By the nature of this method (similarly to the more complex methods sometimes used in modern meta-analysis), one finding with very strong evidence can outweigh multiple null effects. Though Stouffer's Z is sometimes used in meta-analysis, readers are cautioned against interpreting our study as a true meta-analysis – our study involves re-analysis of several datasets that we were able to obtain rather than a traditional meta-analysis, which functions solely from the information available in published papers and attempts to exhaustively survey all relevant papers.

6 RESULTS

6.1 Analyses of Individual Datasets

Table 8 summarizes the results of the individual tests conducted on the 12 non-self-transitions in the 12 affect datasets, with a corrected *L* metric [50]. Across the possible 140 results (4 transitions had undefined *L* value), only 24 tests yielded transitions significantly more likely than chance, as compared to 59 tests that resulted in transitions significantly less likely than chance and 57 null results.

Of the 24 tests with significantly positive results, 15 belong to transitions in or out of *engaged concentration*. Correspondingly, the transitions out of *engaged concentration* have relatively few null results. In contrast, transitions out of frustration have the highest number of null results (25) and only one positive result. It is worth noting that, across the dataset overall, *engaged concentration* is the most common affective state, whereas *frustration* is most rare.

Across datasets, some studies seem to have more null values (e.g., vMedic, Crystal Island) than the others (particularly studies involving Physics Playground, ASSISTments). This may be an attribute of these systems, but it also may be due to the quantity of data. The studies with more null results were also the studies with smaller sample sizes or briefer duration of observations.

Looking across the 12 datasets, we find that many of the transitions postulated by [5] are not statistically significantly more likely than chance (and in fact are often less likely than chance), a finding noted in several of those earlier papers (see discussion in section 2). In fact, the only transition where a significant positive result is seen in a major of datasets is *engaged concentration -> confusion*. By contrast, other transitions are almost never statistically more likely than chance: 1/12 datasets for *confusion -> engaged concentration*, 1/12 datasets for *frustration -> confusion,* and 0/12 datasets for *frustration -> boredom*. It is also true that across all possible transitions (including the ones not in the theoretical model), no transition other than *engaged concentration -> confusion* has a majority of positive transitions.

TABLE 8
SIGNIFICANCE OF THE TRANSITIONS TESTED IN THE CURRENT ANALYSIS FOR THE TWELVE AFFECT DATASETS

Dataset#	Learning System / Classroom	ENG_CON	ENG_FRU	ENG_BOR	CON_ENG	CON_FRU	CON_BOR	FRU_ENG	FRU_CON	FRU_BOR	BOR_ENG	BOR_CON	BOR_FRU
1	ASSISTments	+	+	+	+	+	-	Ø	Ø	Ø	+	-	+
2	Classroom	+	+	+	Ø		Ø	Ø	-	Ø	+	-	-
3	Classroom	+	-	-	Ø	-	Ø				-	+	+
4	vMedic	Ø	-	Ø	Ø	Ø	-	Ø	Ø	Ø	Ø	Ø	Ø
5	Crystal Island	-	-	-	Ø	Ø	Ø	-	Ø	-	Ø	Ø	Ø
6	Aplusix	+	+	-	Ø	-	-	Ø	Ø	-	-	Ø	Ø
7A	Scooter Control	-	-	+	-	-	-	Ø	Ø	Ø	Ø	-	-
7B	Scooter Experiment	-	-	-	-	+	-	Ø	Ø	Ø	Ø	-	-
8	SQL-Tutor	Ø	-	-	Ø	-	-	Ø	Ø	Ø	Ø	-	-
9	Physics Playground	-	-	-	-	+	-	-	-	+	-	-	+
10A	Ecolab Control	+	-	-	Ø	-	Ø	-	Ø	Ø	-	Ø	Ø
10B	Ecolab Experiment	+	-	-	Ø	Ø	+	Ø	Ø	Ø	-	-	Ø
	Total +	6	3	3	1	3	1	0	0	1	2	1	3
	Total Ø	2	0	1	8	3	4	8	9	8	5	4	5
	Total -	4	9	8	3	5	7	3	2	2	5	7	4

+ indicates a significant positive transition, - indicates a significant negative transition and Ø indicates a null effect. Transitions never seen or with undefined L value are left blank. A transition from affect1 to affect2 is denoted as "affect1_affect2." For instance, CON_BOR is a transition from confusion to boredom. Transitions hypothesized in the D'Mello & Graesser's model are highlighted in grey.

6.2 Analyses Across Datasets

Since it is possible that the results of individual studies do not provide a sufficient sample to find a significant effect, it is important to test whether or not any of these transitions may be significant if the studies were aggregated. However, as Table 9 shows, aggregating across studies using Stouffer's Z does not increase the number of transitions that show statistically significant, positive effects. Instead, the only transition that is statistically significantly more likely than chance remains *engaged concentration -> confusion*. There are seven other transitions with a statistically significant result, but all of those have a negative Z score, indicating that they are statistically significantly less likely than chance.

Among the other six transitions postulated in the [5] model, only one transition is significantly more likely than chance (*engaged concentration -> confusion*). Two of their transitions have a null result - *confusion -> engaged concentration* and *frustration -> boredom*. Lastly, three transitions in the hypothesized model are significantly less likely than chance - *confusion -> frustration, frustration -> confusion,* and *boredom-> frustration*. It is worth noting that the original

studies in [5] also had little to no support for the transitions *frustration→confusion* and *frustration→boredom*.

TABLE 9
STOUFFER'S Z AND COMBINED P-VALUES FOR THE TWELVE NON-SELF-TRANSITIONS STUDIED IN THIS PAPER.

Transition	Stouffer's Z	Combined p
ENG_CON	**6.770**	**1.28e-11**
ENG_FRU	-10.878	1.46e-27
ENG_BOR	-12.296	9.40e-35
CON_ENG	-1.605	0.108
CON_FRU	-4.863	1.15e-06
CON_BOR	-7.763	8.25e-15
FRU_ENG	-3.906	9.35e-05
FRU_CON	-2.075	0.037
FRU_BOR	-0.007	0.99
BOR_ENG	-1.344	0.178
BOR_CON	-8.885	6.37e-19
BOR_FRU	-3.861	1.12e-04

The transitions significantly more likely than chance are highlighted in bold. A transition from affect1 to affect2 is denoted as "affect1_affect2." For instance, CON_BOR is a transition from confusion to boredom. Transitions hypothesized in the D'Mello & Graesser's model are highlighted in grey.

6.3 Comparison of Datasets from the US and the Philippines

In the next analysis, we investigated if the model is more correct if we restrict the scope of its applicability. In the 12 datasets analyzed in this paper, datasets #1 to #5 were collected in the United States (US), the country where D'Mello's work was conducted, and datasets #6 to #10 were collected in the Philippines. As noted in section 2.1, the manifestation of affect may differ in different cultures. Thus, we analyze whether there is a difference in the significance pattern for the two countries, looking in particular at whether one of the countries conforms better to the theoretical model. Table 10 and Table 11 present the results of these tests for the US and the Philippines group, respectively.

TABLE 10
STOUFFER'S Z AND COMBINED P-VALUES FOR THE DATA COLLECTED IN THE UNITED STATES.

Transition	Stouffer's Z	Combined p
ENG_CON	**18.337**	**4.14e-75**
ENG_FRU	**8.114**	**4.90e-16**
ENG_BOR	-0.511	0.609
CON_ENG	**2.028**	**0.042**
CON_FRU	-2.581	0.009
CON_BOR	-2.183	0.029
FRU_ENG	-1.768	0.077
FRU_CON	0.752	0.452
FRU_BOR	-0.905	0.365
BOR_ENG	**2.110**	**0.034**
BOR_CON	-5.150	2.60e-07
BOR_FRU	0.264	0.791

The transitions significantly more likely than chance are highlighted in bold. A transition from affect1 to affect2 is denoted as "affect1_affect2." For instance, CON_BOR is a transition from confusion to boredom. Transitions hypothesized in the D'Mello & Graesser's model are highlighted in grey.

On combining the p-values from the datasets collected only in the US (Table 10), we see that a greater number of transitions are significantly more likely than chance as compared to the results in Table 9 (the analysis in both countries). Note that all 4 of these are either from *engaged concentration*, the most frequent state in all datasets (*engaged concentration -> confusion; engaged concentration -> frustration*) or into it (*confusion -> engaged concentration; boredom -> engaged concentration*). Yet only two of these transitions (*engaged concentration -> confusion; confusion -> engaged concentration*) belong to the theoretical model [5].

TABLE 11
STOUFFER'S Z AND COMBINED P-VALUES FOR THE DATA COLLECTED IN THE PHILIPPINES.

Transition	Stouffer's Z	Combined p
ENG_CON	-6.634	3.26e-11
ENG_FRU	-21.100	7.88e-99
ENG_BOR	-15.668	2.46e-55
CON_ENG	-3.816	1.35e-04
CON_FRU	-4.144	3.40e-05
CON_BOR	-8.319	8.80e-17
FRU_ENG	-3.561	3.69e-04
FRU_CON	-2.973	0.0029
FRU_BOR	0.674	0.499
BOR_ENG	-3.543	3.94e-04
BOR_CON	-7.281	3.32e-13
BOR_FRU	-5.279	1.29e-07

The transitions significantly more likely than chance are highlighted in bold. A transition from affect1 to affect2 is denoted as "affect1_affect2." For instance, CON_BOR is a transition from confusion to boredom. Transitions hypothesized in the D'Mello & Graesser's model are highlighted in grey.

In contrast, none of the transitions are significantly more likely than chance when the p-values in the datasets from the Philippines is combined (Table 11).

Hence, affect transitions appear to be more stable in the United States than in the Philippines, but neither country shows patterns that conform particularly well to the theoretical model.

7 DISCUSSIONS

D'Mello and Graesser's model [5] has been one of the most influential theoretical frameworks in affect dynamics research. It theorizes how affect develops over time during learning and describes how the transitions in affect that are hypothesized may contribute to processes of learning and disengagement.

Despite this model's influence on the research community, our survey of the published literature in this area indicates that most of the empirical studies on affect dynamics do not conform to the theoretical model. Even the two empirical studies presented in support of the model in the original paper [5] do not fully support all the hypothesized transitions.

Further investigation of the literature reveals that at least some of the differences in the literature may be culturally driven. The studies that do show some evidence for the model

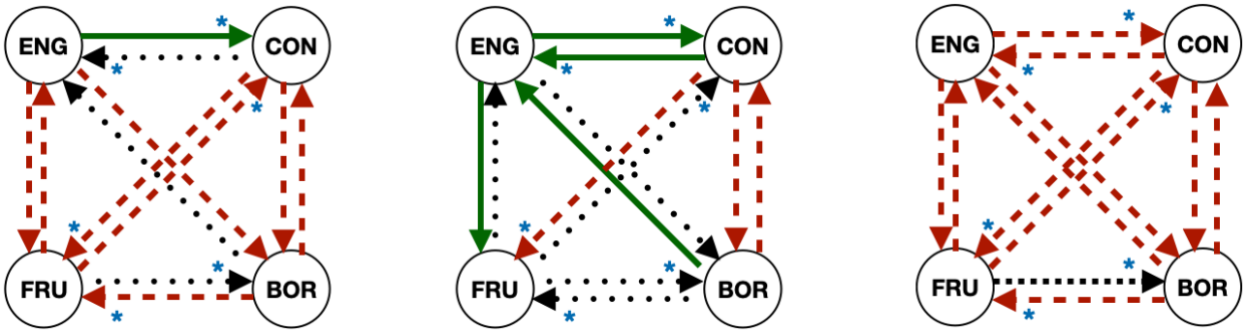


Fig. 3. Visualization of the significantly likely (green solid arrows), significantly unlikely (red dashed arrows), and null (black dotted arrows) transitions using combine p-values from – (a) all the datastes combined, (b) data collected in the US, and (c) data collected in the Philippines. Transitions hypothesized in the D'Mello and Graesser's model is marked with a blue * next to the arrowhead.

were all conducted in the United States with undergraduate populations, but other student populations seem to show more variance in their transition patterns.

In this paper, we reanalyzed and synthesized the data collected in ten publications (twelve datasets) from diverse learning contexts. Our goal was to better understand the pattern of results across these datasets and see what empirical evidence we find for D'Mello and Graesser's theoretical model.

Two methodological concerns drove our decision to reanalyze the data from past publications rather than just synthesizing their results. First, it was not clear from any of the past articles on affect dynamics how certain edge cases were handled, possibly impacting their results. Second, upon investigating methodological differences between the past studies, we observed that the studies that were showing some evidence for the model (including D'Mello and Graesser's studies) incorporated a pre-processing technique (removing self-transitions) that can sometimes produce spurious false positives.

To address these concerns, we first presented a detailed description of the steps involved in affect dynamics analysis, clarifying the edge cases. We also proposed a new correction to the interpretation of the transition metric for reanalysis. By further investigating which factors are associated with studies matching the predictions in D'Mello and Graesser's model (i.e., studies in different countries), we seek to better understand not just its validity but its scope of applicability.

7.1 Non-Conformance to the Theoretical Model

By reanalyzing 12 datasets using D'Mello and Graesser's approach (but with the correction to the transition metric), we show that the data generally does not seem to back up the D'Mello and Graesser [5] model (Figure 3). When all data is analyzed using Stouffer's Z, only one (*engaged concentration -> confusion*) of the six hypothesized transitions is more likely than chance, and no other transitions are more likely than chance (Table 9). This finding indicates that the differences between D'Mello and Graesser's hypothesized model and past published results are not simply due to differences in the analytical method. At best, we can conclude that this widely-accepted model of affect dynamics has a more limited scope than what it is currently being used for. The future use of this model needs a thorough exploration of other aspects of design or contexts to understand where it is an accurate depiction of affective processes.

It is worth noting that most of D'Mello and colleagues' data were collected in lab settings, while the other datasets are from real-world classrooms. Is it possible that real-world events cause an affect to shift more rapidly than in the lab? Is it possible that the coarser grain size of BROMP as compared to retroactive affect judgements is part of what gives us a different result? One way to investigate this question in real-world learning is to use affect detectors to get a finer-grained look at affective processes (i.e. [4]), but given that only one study has used this method so far, and that paper appears to have made the same error around the transition metric as seen in several other studies, more work awaits the synthesis of findings using this method.

There has been considerable interest over the last few years in better understanding the dynamics and trends of affect. The primary assumption here is that learners do not randomly shift between emotions and that there are systematic, recurrent shifts between certain states during learning. Our results suggest that affect may instead be generally irregular, raising the question of whether modeling affect dynamics, *in general,* is still fruitful or useful. Our results suggest that it is highly unlikely that there is a general multi-step pattern in affect dynamics like the *engaged concentration -> confusion -> frustration -> boredom* trend suggested by the theoretical model. However, there may still be some contextually relevant patterns useful to understand the student experience. For instance, students in US classrooms (Table 10) may oscillate back and forth between *engaged concentration* and other states. But more broadly, perhaps we should be looking at the affective changes associated with specific events during student learning more than overall trends and patterns.

7.2 Methodological Implications for Future Affect Dynamics Research

For around a decade, affect dynamics researchers have used the metric L to evaluate the probability of transitions in affect. L is largely believed to have a value of 0 when a transition is at chance, and this is true for the original use of the metric. However, this study provides mathematical evidence that the exclusion of self-transitions leads to a violation of the assumption of independence in the equations used to calculate L. Therefore, this metric does not have a value of 0

at chance if self-transitions are removed.

The primary implication of this finding is in how the L value is interpreted to understand the direction of a transition. If an affective dynamics study excludes self-transitions, we find that when self-transitions are excluded, the value for L that represents chance shifts from 0 to $1/(n-1)^2$, where n is the number of affective states studied. Accordingly, the test for the significance of these transitions must be adjusted so that the null hypothesis is set at the appropriate chance levels and not zero.

This finding, thus, has important implications for the interpretation of past publications. For instance, for a study with four affective states, transitions with an L value less than 0.11 should be interpreted as being less likely than chance. In past studies that excluded self-transitions [4, 5, 21, 23, 37, 38, 50] like the original paper from [5], results must be reinterpreted in terms of the corrected chance value.

In cases where we were unable to analyze the raw data, results need to be reinterpreted on the basis of appropriate chance values for L, given in Table 5. For instance, in first study presented in [5], the transition *confusion -> frustration* is reported to have an L = 0.060 and is significant with $p < 0.05$. This is interpreted as a transition that is more likely than chance, but since self-transitions were removed, researchers should apply a corrected chance value to L (L = 0.11, as shown for n=4 in Table 5) before interpreting this result. This means that the *confusion -> frustration* transition is actually less likely than chance, and the same is true for six of the other ten statistically significant transitions in their two studies.

At the same time, it is important to remind the reader that many past publications using L are unaffected by this concern. Over half of the past studies using this metric included self-transitions [3, 19, 24, 25, 26, 27, 28, 29, 30, 44, 51] and are therefore unchanged by this finding. The choice of whether or not one ought to include self-transitions in an affect dynamics analysis depends on the research goals and questions of the study. Excluding self-transitions reveals a larger number of affective patterns that might otherwise be suppressed by the presence of persistent affective states (although, as our findings indicate, relatively few of these patterns appear to be consistent across studies). Including self-transitions in analysis helps us to better understand each state's persistence, but dilutes the transitions between different affective states. Better understanding transitions is likely important in theoretical models, but understanding persistence might be particularly useful for algorithms being used to trigger interventions, for example.

Recent research has also focused on finding alternative approaches to conducting affect dynamics research, including an update to the L statistic formula [59, 60], and use of epistemic network analysis [61], and marginal models [62].

7.3 Need to Focus on Cultural Factors in Affect Dynamics Research

One other important finding in this study is that affective patterns seem to differ based on the country in which the research was conducted (US versus Philippines). Across studies, no affective transitions were more likely than chance in the Philippines, while there were 4 significant transitions in the US (*engaged concentration -> confusion; engaged concentration -> frustration*; *confusion -> engaged concentration; boredom -> engaged concentration*). A similar pattern can be seen in past affect dynamics studies from the Philippines (Table 3), which included self-transitions (our analysis excluded them).

Currently, it is not clear why affect dynamic results are so different in the Philippines and the United States. It is not that the most common affect differs – this seems to be relatively consistent across studies, many of which started with additional affective states. It is not that the relationship between affect and learning differs – the negative correlation between boredom and learning and the positive correlation between *engaged concentration* and learning are seen in both countries (along with the instability of correlation within each country for confusion and frustration) (i.e., 43, 52, 63, 64). Despite these commonalities, the affect dynamics seem to differ.

Given the many differences between schools in the United States and the Philippines – national culture, school culture, use of educational technology, prevalent forms of disengagement [65] – it is difficult at this point to understand *why* we see these differences. It may even be the case that the affect being recognized in different countries is fundamentally different in kind, in a way that the researchers conducting these studies cannot fully recognize. The BROMP field observation protocol was co-developed by researchers in the USA and Phillipines and has now been applied in several other countries, but that does not guarantee that the same constructs are captured when a researcher in each country identifies "engaged concentration" or "frustration". Indeed, many individuals have found it difficult to achieve acceptable inter-rater reliability out of their native culture [44], and there are systematic biases in cross-cultural attempts to recognize affect [64]. A better understanding of the role that culture plays in the manifestation and recognition of affect is important for any future attempts to study affect dynamics as a generalizable phenomenon. Ultimately, it may make the best sense to re-consider that question after affect dynamics has been studied in a wider range of cultures, potentially looking at the kinds of traits that are known to vary at a national level.

7.4 Limitations and Future Work

There are some limitations to this paper's findings. First, all but one dataset among the ten studies are collected through quantitative field observations, which may sample at slower rates than many other approaches to collect affect data, like video, sensors, emote-aloud methods, and self-reports. This choice was driven primarily by the previous papers that investigate the phenomena of interest, as well as the availability of datasets. However, other methods have different virtues and flaws in terms of cost, scale, accuracy, and time. It would be interesting for future work to explore how each of these methods captures student emotions differently and how these differences impact the validity or applicability of the affect dynamics analysis.

Second, this analysis did not consider the impact of other attributes of the study design like observation grain size and the length of observation session. As we note above, finer-grained affect observations could potentially yield a different result than the coarser-grained data from BROMP observations. Moreover, it seems likely that students may be more likely to hit points of frustration and boredom later rather

than earlier in a long observation system, particularly if they were working within a learning system that became increasingly challenging over time.

Third, all our significant transitions have *engaged concentration* in them. *Engaged concentration* also happens to be the most frequent affective state in all our datasets. In contrast, *frustration* is the rarest affective state and has the most null or negative results. Thus, future work needs to analyze if there exists a threshold for the minimum length of the affect sequence or minimum base rate of each affective state to be able to see significant positive transitions. Alternatively, it may be worth explicitly seeking out more difficult tasks and contexts where frustration may be more common, such as learning systems that are known to be less effective or students working alone at home within fully asynchronous virtual schooling. It is also possible that studies have seen less frustration because most are conducted in short-term studies rather than ongoing use; studies like [38] that involve an entire year of usage may be more able to detect affective sequences associated with frustration.

Fourth, this work synthesizes across multiple affect datasets. Like the studies it builds on, it does not consider individual differences in affect incidence and dynamics that may appear in the data. It is possible that different patterns – both related and unrelated to the theoretical model – may be characteristic of sub-groups. Differences in affective dynamics may be associated with a variety of individual differences, such as differences in personality.

Overall, this paper provides a comprehensive look at affect dynamics across published work. Broadly, work so far does not seem to accord with the most popular theoretical model. Further work is needed to understand what is general in the dynamics of affect, both for specific contexts and across contexts.

ACKNOWLEDGMENT

Our thanks to Penn Center for Learning Analytics for funding this research. The authors would like to thank Dr. Nigel Bosch, Dr. Luc Paquette and to Dr. Anthony Botelho for their early inputs while developing this work. The authors would further like to express their gratitude to Dr. Anna Fisher, Dr. Anthony Botelho, Dr. Douglas DiStefano, Dr. James Lester, Dr. Jennifer Sabourin, Juan Miguel Andres-Bray, Dr. Karrie E. Godwin, Dr. Ma. Mercedes Rodrigo, Dr. Scott McQuiggan and Dr. Thea Faye Guia for sharing the affect datasets used in this anlaysis.

REFERENCES

[1] McQuiggan, S. W., & Lester, J. (2009). Modeling affect expression and recognition in an interactive learning environment. *International Journal of Learning Technology, 4*(3-4), 216-233.

[2] D'Mello, S., Person, N., Lehman, B. (2009). Antecedent-Consequent Relationships and Cyclical Patterns between Affective States and Problem Solving Outcomes. In *AIED*, 57-64.

[3] Rodrigo, M.M.T., Anglo, E., Sugay, J., Baker, R. (2008). Use of un-supervised clustering to characterize learner behaviors and affective states while using an intelligent tutoring system. In *International Con-ference on Computers in Education*, 57-64.

[4] Bosch, N., & D'Mello, S. (2017). The Affective Experience of Novice Computer Programmers. *International Journal of Artificial Intelligence in Education*, 1-26.

[5] D'Mello, S. Graesser, A. (2012). Dynamics of Affective States during Complex Learning. *Learning and Instruction*, 22, 145-157.

[6] Arroyo, I., Cooper, D. G., Burleson, W., Woolf, B. P., Muldner, K., & Christopherson, R. (2009, July). Emotion sensors go to school. In *AIED* (Vol. 200, pp. 17-24).

[7] Jaques, N., Conati, C., Harley, J. M., & Azevedo, R. (2014, June). Predicting affect from gaze data during interaction with an intelligent tutoring system. In *International conference on intelligent tutoring systems* (pp. 29-38). Springer, Cham.

[8] D'Mello, S., & Kory, J. (2012, October). Consistent but modest: a meta-analysis on unimodal and multimodal affect detection accuracies from 30 studies. In *Proceedings of the 14th ACM international conference on Multimodal interaction* (pp. 31-38).

[9] Nye, B. D., Karumbaiah, S., Tokel, S. T., Core, M. G., Stratou, G., Auerbach, D., & Georgila, K. (2018, June). Engaging with the scenario: Affect and facial patterns from a scenario-based intelligent tutoring system. In *International Conference on Artificial Intelligence in Education* (pp. 352-366). Springer, Cham.

[10] Cambria, E., Das, D., Bandyopadhyay, S., & Feraco, A. (2017). Affective computing and sentiment analysis. In *A practical guide to sentiment analysis* (pp. 1-10). Springer, Cham.

[11] Cambria, E., Li, Y., Xing, F. Z., Poria, S., & Kwok, K. (2020, October). SenticNet 6: Ensemble application of symbolic and subsymbolic AI for sentiment analysis. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management* (pp. 105-114).

[12] DeFalco, J.A., Rowe, J.P., Paquette, L., Georgoulas-Sherry, V., Brawner, K., Mott, B.W., Baker, R.S., Lester, J.C. (2018) Detecting and Addressing Frustration in a Serious Game for Military Training. *International Journal of Artificial Intelligence and Education, 28* (2), 152-193.

[13] Karumbaiah, S., Lan, A., Nagpal, S., Baker, R. S., Botelho, A., & Heffernan, N. (2021, April). Using Past Data to Warm Start Active Machine Learning: Does Context Matter?. In *LAK21: 11th International Learning Analytics and Knowledge Conference* (pp. 151-160).

[14] Sanghvi, J., Castellano, G., Leite, I., Pereira, A., McOwan, P. W., & Paiva, A. (2011, March). Automatic analysis of affective postures and body motion to detect engagement with a game companion. *In Proceedings of the 6th International Conference on Human-robot Interaction* (pp. 305-312). ACM.

[15] Clavel, C., & Callejas, Z. (2015). Sentiment analysis: from opinion mining to human-agent interaction. *IEEE Transactions on affective computing, 7*(1), 74-93.

[16] DeFalco, J. A., Rowe, J. P., Paquette, L., Georgoulas-Sherry, V., Brawner, K., Mott, B. W., Baker, R. S., & Lester, J. C. (2018). De-tecting and addressing frustration in a serious game for military train-ing. *International Journal of Artificial Intelligence in Education, 28* (2), 152-193.

[17] Karumbaiah, S., Lizarralde, R., Allessio, D., Woolf, B. P., Arroyo, I., & Wixon, N. (2017). Addressing Student Behavior and Affect with Empathy and Growth Mindset. *Proceedings of the 10th International Conference on Educational Data Mining*.

[18] Kuppens, P. (2015). It's about time: A special section on affect dynamics. *Emotion Rev., 7*(4), 297-300.

[19] Andres, J.M.L., & Rodrigo, M.M.T. (2014). The Incidence and persistence of affective states while playing Newton's playground. *7th IEEE International Conference on Humanoid, Nanotechnology, Information Tech., Communication and Control, Environment, and Management.*

[20] Baker, R.S., Rodrigo, M.M.T., Xolocotzin, U. (2007). The dynamics of affective transitions in simulation problem-solving environments. *International Conf. on Affective Computing and Intelligent Interaction. Springer Berlin Heidelberg,* 666-67.

[21] Bosch, N., & D'Mello, S. (2013). Sequential patterns of affective states of novice programmers. In *The 1st Workshop on AI-supported Education for Computer Science (AIEDCS 2013)*, 1-10.

[22] D'Mello, S., Graesser, A. (2015). Feeling, thinking, & computing with affect-aware learning technologies. Calvo, D'Mello, Gratch, Kappas (eds.) *Handbook of Affective Computing*. Oxford UP.

[23] D'Mello, S., Taylor, R., & Graesser, A. (2007). Monitoring Affective Trajectories during Complex Learning. D. McNamara & J. Trafton (Eds.), *Proc. 29th Annual Cognitive Science Soc.*, 203-8.

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

[25] Guia, T.F.G., Sugay, J., Rodrigo, M.M.T., Macam, F., Dagami, M., Mitrovic, A. (2011). Transitions of Affective States in an Intelligent Tutoring System. *Proc. Philippine Computing Soc.* 31-5.

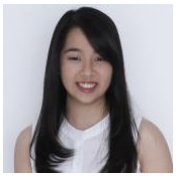
[26] McQuiggan, S. W., Robison, J. L., & Lester, J. C. (2010). Affective transitions in narrativecentered learning environments. *Educational Technology & Society, 13*(1), 40-53.

[27] Mitrovic, A. (2003). An intelligent SQL tutor on the web. *International Journal of Artificial Intelligence in Education, 13(*2-4), 173-197.
[28] Ocumpaugh, J., Andres, J.M., Baker, R., DeFalco, J., Paquette, L., Rowe, J., et al. (2017). Affect Dynamics in Military Trainees using vMedic: From Engaged Concentration to Boredom to Confusion. In *International Conf. on Artificial Intelligence in Ed*., 238-249. Springer, Cham.
[29] Rodrigo, M.M.T., Baker, R., Agapito, J., Nabos, J., Repalam, M., Reyes Jr, S., San Pedro, M.O.C. (2011). The effects of an embodied conversational agent on student affective dynamics while using an intelligent tutoring system. *IEEE Transactions on Affective Computing, 2*(4), 18-37.
[30] Rodrigo, M.M.T., Baker, R., Agapito, J., Nabos, J., Repalam, M., Reyes, S., San Pedro, M.O.C. (2012). The effects of an interactive software agent on student affective dynamics while using; an intelli-gent tutoring system. *IEEE Transactions on Affective Computing, 3*(2), 224-36.
[31] Schwarz, N. (2012). Feelings-as-Information Theory. In P. Van Lange, A. Kruglanski & T. Higgins (Eds.), *Handbook of Theories of Social Psychology* (pp. 289-308). Thousand Oaks, CA: Sage.
[32] Izard, C. (2010). The many meanings/aspects of emotion: Definitions, functions, activation, and regulation. *Emotion Review, 2*(4), 363-370.
[33] Barth, C. M., & Funke, J. (2010). Negative affective environments improve complex solving performance. *Cognition and Emotion, 24*(7), 1259-1268.
[34] Fredrickson, B. L., & Branigan, C. (2005). Positive emotions broaden the scope of attention and thought- action repertoires. *Cognition & Emotion, 19*(3), 313-332.
[35] Csikszentmihalyi, M. (1990). *Flow and the psychology of discovery and invention*. Harper Collins, New York.
[36] Baker, R.S., D'Mello, S., Rodrigo, M.M.T., Graesser, A. (2010). Better to be frustrated than bored: The incidence, persistence, & impact of learners' cognitive-affective states during interactions with 3 different computer-based learning environments. *Int'l J. Hum-Comp. Stu., 68* (4),223-41.
[37] D'Mello, S., Graesser, A. (2010). Modeling cognitive-affective dy-namics with Hidden Markov Models. *Proceedings of the 32nd Annual Cognitive Science Society*, 2721-2726.
[38] Botelho, A.F., Baker, R., Ocumpaugh, J., Heffernan, N. (2018) Study-ing Affect Dynamics and Chronometry Using Sensor-Free Detectors. *Proceedings of the 11th International Conference on Educational Data Mining*, 157-166.
[39] Tsai, J., Levenson, R. (1997). Cultural influences on emotional re-sponding: Chinese Am. & European Am. dating couples during inter-personal conflict. *J. Cross-Cultural Psych., 28*(5), 600-25.
[40] Kitayama, S., Markus, H. R., & Kurokawa, M. (2000). Culture, emo-tion, and well-being: Good feelings in Japan and the United States. *Cognition & Emotion, 14*(1), 93-124.
[41] Dunn, J., & Brown, J. (1994). Affect expression in the family, chil-dren's understanding of emotions, and their interactions with others. *Merrill-Palmer Quarterly* (1982-), 120-137.
[42] Gross, J. J., Carstensen, L. L., Pasupathi, M., Tsai, J., Götestam Skorpen, C., & Hsu, A. Y. (1997). Emotion and aging: Experience, expression, and control. *Psychology and Aging, 12*(4), 590.
[43] 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.
[44] Baker, R.S., Ocumpaugh, J.L., Andres, J.M.A.L. (in press) BROMP Quantitative Field Observations: A Review. In R. Feldman (Ed.) Learning Science*: Theory, Research, and Practice*. New York, NY: McGraw-Hill.
[45] Ocumpaugh, J., Baker, R.S., Rodrigo, M.M.T. (2015). *Baker Rodrigo Ocumpaugh Monitoring Protocol (BROMP) 2.0 Technical & Training Manual.* Technical Report. NY, NY: Teachers College, Columbia U. Manila, Philippines: Ateneo Laboratory for the Learning Sciences.
[46] Gonzalez, C., Best, B., Healy, A., Kole, J. A., Bourne Jr, L. (2011). A cognitive modeling account of simultaneous learning and fatigue ef-fects. *Cognitive Systems Research, 12*(1), 19-32.
[47] Healy, A., Kole, J., Buck-Gengler, C., Bourne Jr, L., (2004). Effects of prolonged work on data entry speed and accuracy. *Journal of Experimental Psychology: Applied,* 10, 188–199.
[48] Karumbaiah, S., Andres, J. M. A. L., Botelho, A. F., Baker, R. S., & Ocumpaugh, J. S. (2018). The Implications of a Subtle Difference in the Calculation of Affect Dynamics. *In 26th International Conference for Computers in Education.*
[49] Andres, J.M.L., Rodrigo, M.M T., Sugay, J., Banawan, M., Paredes, Y., Cruz, J., Palaoag, T. (2015). More Fun in the Philippines? Factors Affecting Transfer of Western Field Methods to One Developing World Context. In *AIED Workshops.*
[50] Karumbaiah, S., Baker, R. S., & Ocumpaugh, J. (2019, June). The Case of Self-transitions in Affective Dynamics. *In International Con-ference on Artificial Intelligence in Education* (pp. 172-181). Springer, Cham.
[51] Rodrigo, M.M.T., Rebolledo-Mendez, G., Baker, R., du Boulay, B., Sugay, J.., Lim, S., Luckin, R. (2008). The effects of motivational modeling on affect in an intelligent tutoring system. *Proc. of International Conference on Computers in Education*, 57, 64.
[52] Rodrigo, M. M. T., & Baker, R. S. (2009, August). Coarse-grained detection of student frustration in an introductory programming course. In *Proceedings of the 5th International Workshop on Computing Education Research Workshop* (pp. 75-80). ACM.
[53] Godwin, K. E., Almeda, M. V., Seltman, H., Kai, S., Skerbetz, M. D., Baker, R. S., & Fisher, A. V. (2016). Off-task behavior in elementary school children. *Learning and Instruction*, 44, 128-143.
[54] DiStefano, D. (2018*). How Pre-Service Teachers' Engagement and Affect Informs Instructional Format of an Introductory Methods Course* (Doctoral dissertation, Fordham University).
[55] Nicaud, J. F., Bouhineau, D., Mezerette, S., & Andre, N. (2007). *Aplusix II* [Computer software].
[56] Shute, V., Ventura, M. (2013). Stealth assessment: Measuring & supporting learning in video games. *MIT Press*
[57] Stouffer, S.A., Suchman, E.A., DeVinney, L.C., Star, S.A. & Williams, R.M. Jr. (1949). *The American Soldier, Vol. 1: Adjustment During Army Life.* Princeton University Press, Princeton.
[58] Rosenthal, R., & Rosnow, R. L. (1991). *Essentials of behavioral research: Methods and data analysis* (Vol. 2). New York: McGraw-Hill.split
[59] Bosch, N & Paquette, L. (2020). What's Next? Edge Cases in Measuring Transitions Between Sequential States. Submitted
[60] Matayoshi, J., & Karumbaiah, S. (2020). Adjusting the L Statistic when Self-Transitions are Excluded in Affect Dynamics. *Journal of Educational Data Mining, 12*(4), 1-23.
[61] Karumbaiah, S., & Baker, R. S. (2021, February). Studying Affect Dynamics using Epistemic Networks. In *International Conference on Quantitative Ethnography* (pp. 362-374). Springer, Cham.
[62] Matayoshi, J., & Karumbaiah, S. (2021, April). Using Marginal Models to Adjust for Statistical Bias in the Analysis of State Transitions. In *LAK21: 11th International Learning Analytics and Knowledge Conference* (pp. 449-455).
[63] Lagud, M. C. V., & Rodrigo, M. M. T. (2010, June). The affective and learning profiles of stu-dents using an intelligent tutoring system for algebra. In *International Conference on Intelligent Tutoring Systems* (pp. 255-263). Springer, Berlin, Heidelberg.
[64] Okur, E., Aslan, S., Alyuz, N., Esme, A.A., Baker, R.S. (2018) Role of Socio-Cultural Differences in Labeling Students' Affective States. Proceedings of the *19th International Conference on Artificial Intelligence in Education*, 367-380.
[65] Rodrigo, M. M. T., Baker, R. S. J. D., & Rossi, L. (2013). Student off-task behavior in computer-based learning in the Philippines: comparison to prior research in the USA. *Teachers College Record, 115*(10), 1-27.
[66] Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and psychological measurement, 20*(1), 37-46.

Shamya Karumbaiah received the bachelor's degree in computer science in 2011 from Sri Jayachamarajendra College of Engineering, India. In 2017, she earned a M.S in computer science from the University of Massachusetts Amherst with an emphasis in machine learning. She is currently pursuing her PhD in learning sciences from the University of Pennsylvania. She worked as a software engineer at Cisco Systems between 2011-15. In 2016, she was a visiting researcher at the University of Southern California Institute for Creative Technologies. She was a data science intern at the advanced technologies and AI lab at Cisco Systems in 2017. Four of her first-authored papers were nominated for the best overall paper in LAK 2021, ICQE 2020, EDM 2018, and ICCE 2018. Her current research interests include promoting student engagement and learning in virtual learning environments in a fair and equitable manner.

Ryan S. Baker received his Sc.B. from Brown University in 2000 and his Ph.D. in Human-Computer Interaction from Carnegie Mellon University in 2005. He was faculty at Worcester Polytechnic Institute and Teachers College, Columbia University, and is now Associate Professor at the University of Pennsylvania. He is editor of Computer-Based Learning in Context. He has written over 350 papers, 14 of which have received conference paper awards, and has won the Educational Research Award from the Council of Scientific Society Presidents. He is interested in educational data mining and predictive analytics.

Jaclyn Ocumpaugh has a BA in Linguistics from The University of Texas, Austin (1999), an MA in English Language and Linguistics from North Carolina State University, Raleigh (2002), and a PhD in Linguistics from Michigan State University, East Lansing (2010). She has worked for Worcester Polytechnic Institute; Teachers College, Columbia University; and now the University of Pennsylvania. She and her team have been nominated for and/or won over a half a dozen best paper across over 60 publications to date. She currently specializes in student engagement and affective research.



Juliana Ma. Alexandra L. Andres earned her B.A. in Psychology at the Ateneo de Manila University in 2017 and earned her M.S.Ed. from the University of Pennsylvania in 2018. She is currently pursuing her Ed.D. at the University of Pennsylvania. She worked as a research assistant in the Ateneo Laboratory for the Learning Sciences and the Political Psychology Laboratory between 2015 and 2017 before joining the Penn Center for Learning Analytics. Andres has co-authored two conference full papers that have been nominated for best paper awards, one of which had won best student paper. She has currently published five conference full papers, one journal paper, and one book chapter. Her current research interests involve student affect and behavior within intelligent tutoring systems.