Int J Artif Intell Educ
DOI 10.1007/s40593-017-0152-1

ARTICLE

Detecting and Addressing Frustration in a Serious Game for Military Training

Jeanine A. DeFalco[1] · **Jonathan P. Rowe**[2] · **Luc Paquette**[3] · **Vasiliki Georgoulas-Sherry**[1] · **Keith Brawner**[4] · **Bradford W. Mott**[2] · **Ryan S. Baker**[5] · **James C. Lester**[2]

Abstract Tutoring systems that are sensitive to affect show considerable promise for enhancing student learning experiences. Creating successful affective responses requires considerable effort both to detect student affect and to design appropriate responses to affect. Recent work has suggested that affect detection is more effective when both physical sensors and interaction logs are used, and that context-sensitive design of affective feedback is necessary to enhance engagement and improve learning. In this paper, we provide a comprehensive report on a multi-part study that integrates detection, validation, and intervention into a unified approach. This paper examines the creation of both sensor-based and interaction-based detectors of student affect, producing successful detectors of student affect. In addition, it reports results from an investigation of motivational feedback messages designed to address student frustration, and investigates whether linking these interventions to detectors improves outcomes. Our results are mixed, finding that self-efficacy enhancing interventions based on interaction-based affect detectors enhance outcomes in one of two experiments investigating affective interventions. This work is conducted in the context of the GIFT framework for intelligent tutoring, and the TC3Sim game-based simulation that provides training for first responder skills.

✉ Jeanine A. DeFalco
jad2234@tc.columbia.edu

[1] Teachers College, Columbia University, New York, NY, USA

[2] North Carolina State University, Raleigh, NC, USA

[3] University of Illinois at Urbana-Champaign, Champaign, IL, USA

[4] U.S. Army Research Laboratory, Orlando, FL, USA

[5] University of Pennsylvania, Philadelphia, PA, USA

Published online: 12 September 2017

Keywords Affect detection · Motivational feedback · Game-based learning · Gift

Introduction

Affect is critical to understanding learning. However, the interplay between affect and learning is complex. Some affective states, such as boredom, have been shown to be associated with reduced learning outcomes (Craig et al. 2004). Other affective states, such as engaged concentration, have been found to serve beneficial roles (Pardos et al. 2014). Fundamentally, the ability to *detect* a learner's affective state while she interacts with an online learning environment is required for adaptive learning technologies that aim to support and regulate learners' affect through *intervention* (D'Mello et al. 2013; Robison et al. 2009).

Research on affective learning environments has enabled the development of models that automatically detect learner affect using a wide variety of data modalities; an extensive review can be seen in Calvo and D'Mello (2010). Many researchers have focused on physical sensors, because of their capacity to capture physiological and behavioral manifestations of emotion. The potential of the physical sensors is that the technology could be applicable to any learning system. Sensor-based detectors of affect have been developed using a range of physical indicators including facial expressions (Arroyo et al. 2009; Bosch et al. 2015), voice (Zeng et al. 2009), posture (D'Mello and Graesser 2010; Grafsgaard et al. 2012), physiological data (Nasoz et al. 2004), smart phone and wristband accelerometers (Garcia-Ceja et al. 2016; Gjoreski 2016) and EEG (AlZoubi et al. 2009). Despite this promise, deploying physical sensors in real-world settings is challenging (Baker and Ocumpaugh 2015), and efforts in this area are still ongoing, with some researchers arguing that this type of affect detection has not yet reached its full potential (D'Mello and Kory 2015).

In recent years, efforts have also been made towards the development of complementary affect detection techniques that recognize affect solely from logs of learner interactions with an online learning environment (Pardos et al. 2014; Paquette et al. 2014). Initial results in this area have shown considerable promise. As both sensor-based and interaction-based affect detectors continue to mature, efforts are needed to compare the relative advantages of each approach. An early comparison was seen in D'Mello et al. (2008), and a more recent comparison was offered in Kai et al. (2015), but there remains surprisingly little attention to this important area.

In this paper, we report on the findings of a three-year project that included three successive studies. These studies include 1) a baseline study to develop sensor-based and interaction-based detectors for the detection of frustration, 2) a subsequent study to verify these interaction-based affect detectors while evaluating the effects of three distinct motivational intervention feedback messages, and 3) a study that compares interventions based on sensor-based detectors, interaction-based detectors, and yoked interventions given to all students.

Project Overview

In this project, we develop and leverage automated detectors to infer the affect of trainees learning through TC3Sim, a serious game for training tactical combat casualty

care skills. Within this project, we combined interaction-based methods for detecting student frustration (e.g., models making inferences solely from the trainee's interaction history within the game-based simulation) with scalable sensor-based methods for detecting affect, with the eventual aim of developing affect models that can leverage sensor information when it is available, but which can still assess trainee affect even when sensors are not available (e.g. Bosch et al. 2015). The automated detectors were developed and validated for accuracy when applied to new trainees.

The automated detectors were then integrated with the TC3Sim training environment to drive run-time feedback messages for trainee affect. This integration was realized through the Generalized Intelligent Framework for Tutoring (GIFT) (Sottilare et al. 2012), a software framework that includes a suite of tools, methods, and standards for research and development of intelligent tutoring systems and affective learning environments. TC3Sim had previously been integrated into GIFT for use on studies of feedback source modality effects (Goldberg and Cannon-Bowers 2015) and authoring tools (Mall and Goldberg 2014).

Several candidate types of feedback were tested for efficacy on enhancing affect and learning compared to a control condition. The best of these interventions was then used to test whether sensor-based or interaction-based triggers for affective feedback were more effective than simply administering interventions to all students.

TC3Sim Combat Medic Training Simulation

The research discussed here was conducted within the context of TC3Sim, a serious game used to train US Army combat medics and combat lifesavers (i.e., nonmedical soldiers with training on providing lifesaving care) on tasks associated with dispensing tactical combat casualty care, including care under fire (Sotomayor 2010) (see Fig. 1). Also known as vMedic, TC3Sim has been integrated with the Generalized Intelligent Framework for Tutoring (GIFT). The training system, as well as many of the scenarios used throughout this paper, has been made publicly available in an effort to promote further study.[1]

Serious games enable a broad range of problem scenarios for learning (Baker et al. 2012; Pardos et al. 2014). For example, game-based learning environments such as EcoMUVE (Metcalf et al. 2011) and Crystal Island (Rowe et al. 2009) provide learners with considerable agency to explore virtual worlds as they see fit. These open-world games encourage learners to choose their own problem-solving paths and self-regulate learning. They can be contrasted with games that encourage learners to adhere to specific problem-solving sequences, as well as environments that utilize game features to "reward" learner performance. Problem scenarios in TC3Sim are structured linearly, presenting a fixed series of events regardless of the learner's actions. However, learners can freely navigate the 3D virtual environments in TC3Sim, as well as choose how to administer casualty care under fire. While TC3Sim supports a considerable amount of learner agency, its training scenarios are designed to direct players towards the game's objectives (e.g., administering care), guiding the learner toward key learning objectives. Players experience these learning objectives through simple care-based scenarios where an injured Soldier and instructor/tutor are the only items of interest in an otherwise

[1] A downloadable version of TC3Sim is available at: https://www.gifttutoring.org/projects/gift/files

Fig. 1 TC3Sim, a serious game for tactical combat casualty care training

barren clearing, as well as through scenarios where the participant witnesses the wounding of a Soldier who needs care.

Background

Sensor-Free, Interaction-Based Detectors of Learner Engagement and Affect

In recent years, there has been increasing work on developing automated online detectors of students' behaviors associated with engagement and affect, inferring these constructs solely from students' patterns of interaction with educational software. Following early work using data mining methods to infer a student's disengaged "gaming the system" behavior (e.g. Baker et al. 2004), models have been developed which can infer the following behaviors associated with engagement or disengagement: off-task behavior (Baker 2007a, b; Cetintas et al. 2009), self-explanation — an engaged behavior (Shih et al. 2008; Baker et al. 2011), carelessness (San Pedro et al. 2011a, b; Hershkovitz et al. 2011), and WTF ("without thinking fastidiously") behavior, actions within a learning system not targeted towards learning or successful performance (Wixon et al. 2012; cf. Rowe et al. 2009).

In terms of affect, sensor-free models have been developed that can infer confusion (Pardos et al. 2014; Liu et al. 2013), boredom (Baker et al. 2012; D'Mello et al. 2008; Sabourin et al. 2011), frustration (D'Mello et al. 2008; Baker et al. 2012; Liu et al. 2013; Pardos et al. 2014; Paquette et al. 2014), and engaged concentration (Baker et al. 2012; D'Mello et al. 2008; Sabourin et al. 2011). These automated detectors have been used to conduct basic research on the conditions and impacts of engagement and affect on learners, including research on the relationship between these constructs and learning (Baker et al. 2004; Cocea et al. 2009; Pardos et al. 2014), student goal orientation (Baker 2007a, b; Baker et al. 2008; Hershkovitz et al. 2013), and student attitudes towards mathematics (Arroyo et al. 2009; Baker 2007a, b; Baker et al. 2008). Once a detector is developed and validated, it can be applied at scale to additional data, potentially enabling very large-scale analyses (see discussion in Hollands and Bakir 2015).

These automated detectors have been embedded into educational software, identifying situations where automated interventions are appropriate. For example, D'Mello

and Graesser (2010) conducted a study comparing an affect-sensitive version of Auto-Tutor with a non-affective Auto-Tutor. The affect-sensitive Auto-Tutor detected and responded to students' affective states of boredom, confusion, and frustration by monitoring dialogue, body language, and facial features. The results included improved learning for students with comparatively lower domain knowledge who needed more support, than students with comparatively more knowledge. The authors further state that there should be consideration of the appropriate timeliness for providing affect-sensitive feedback to avoid irritating students who have more domain knowledge and do not seem to need affective support in the same way as low-domain knowledge students.

Arroyo et al. (2009) used an interaction-based sensor-free automated detector of gaming to deliver metacognitive messages, also resulting in improved engagement and learning. In a follow up study by Arroyo et al. (2010), three types of intervention messages were used to reduce gaming behavior: attribution interventions, effort-affirmation interventions, and strategic interventions. While these interventions resulted in less gaming the system, less frustration, and more interest as compared to a control condition, there was no impact on learning.

Sensor-Based Detectors of Learner Engagement and Affect

Several research labs have investigated sensor-based affect recognition in learning environments over the past decade, including work with facial expression (Bosch et al. 2015), eye tracking (Jaques et al. 2014), and posture (Cooper et al. 2009; Grafsgaard et al. 2014). In this work, we focus primarily on posture sensor-based models of affect recognition. To date, posture-based affect recognition models have been induced with data from pressure-sensitive chairs (Cooper et al. 2009; D'Mello and Graesser 2010), as well as motion sensors, such as Microsoft Kinect (Grafsgaard et al. 2014). These two data streams, drawing from distinct types of sensors, are superficially different, but can be distilled into analogous predictor features that have similar relationships with affective states such as engagement, boredom, frustration, and confusion. Features can be distilled from both types of data to indicate leaning forward, leaning backward, sitting upright, and fidgeting (D'Mello and Kory 2015).

D'Mello and Graesser (2010) utilized posture data from the Body Pressure Measurement System (BPMS), a type of pressure-sensitive seating device, to predict judgments of student affect during learning with Auto-Tutor. In their study, participants were video recorded, and several judges analyzed the video using freeze frame analysis in order to code participants' affective states retrospectively. Findings indicated that the models, averaged across judges, explained approximately 11% of the variance in students' affective states, with affect states such as delight and flow coinciding with forward leaning, boredom coinciding with a tendency to lean back, and states such as confusion and frustration coinciding with an upright posture.

Cooper et al. (2009) used a suite of sensors to collect data on student affect in Wayang Outpost, an intelligent tutoring system for high school geometry. The sensors included a skin conductance bracelet, pressure sensitive mouse, pressure sensitive chair, and mental state camera, which provided data on student posture, movement, grip tension, arousal, and facial expression. Step-wise linear regression models were induced to predict students' emotion self-reports. Results indicated

that posture-based features were significantly predictive of self-reported excitement during learning, although they were not part of the best-performing models for other emotional states.

Grafsgaard et al. (2014) have investigated postured-based affect prediction using Microsoft Kinect sensors with an intelligent tutoring system for introductory programming. Posture features were distilled from depth image recordings by tracking the distance between the depth camera and the participant's head, upper torso, and lower torso. The posture-based predictor features were combined with features distilled on facial expressions and facial movements to induce multiple regression models for predicting students' retrospective self-reports of engagement and frustration. Findings indicated that posture features were predictive of both self-reported affective states: leaning forward was predictive of both higher engagement and higher frustration, and postural movement was associated with increased frustration and reduced learning.

Building upon this foundation, we set out to distill similar predictor features from the data we collected at the United States Military Academy (USMA) at West Point and apply similar machine learning methods in order to produce affect recognition models for predicting field observations of affect through a series of experiments. As the data from Experiment 1 yielded a negative correlation between the measures of frustration and learning gains, it was decided that the team would focus on how best to intervene with feedback messages within TC3Sim upon the detection of frustration.

Intervening to Address Learner Frustration with Motivational Feedback Messages

Some affective states have relatively uncomplicated relationships with student learning outcomes – for example, engaged concentration appears to be positively associated (Craig et al. 2004; Pardos et al. 2014) while boredom is negatively associated (Craig et al. 2004; Pardos et al. 2014). However, research has shown that the affective state of frustration is more complex, where brief periods of frustration are not problematic, but extended frustration – of, for example, one minute or more (Liu et al. 2013) is associated with worse learning outcomes (D'Mello and Graesser 2011; Liu et al. 2013; Robison et al. 2009). It is important to understand how ITSs can respond to learner frustration for future affect-sensitive learning environments (Picard et al. 2004). Frustration has been related to positive, null, negative, and mixed learning outcomes in ITSs (McQuiggan and Lester 2007 – negative outcomes; Pardos et al. 2014 – positive outcomes; D'Mello et al. 2013 and Rodrigo et al. 2012 – null outcomes; Liu et al. 2013 – mixed outcomes). Given the range and complexity of the impact of frustration on learning outcomes, then, further research was required to unpack the impact of frustration on learning and, equally important, how best to respond to an individual's frustrated state in an intelligent tutoring system; when to do so and how.

When a learner is in a frustrated state in ITSs, the range of solutions to address this frustration includes changing the elements in a system that elicits frustration, and supporting the learner in their ability to recover, manage, and persist in their task (Klein et al. 2002; Kapoor et al. 2007). Amsel's (1992) frustration theory supports the notion that goal attainment includes overcoming emotional conflict rather than avoiding emotional conflict. Therefore, to encourage a learner to overcome frustration, while not changing the nature of the system elements nor providing outright explicit

feedback, requires finding ways to help the learner recover, manage, and persist through frustration to persist in their learning tasks (Kapoor et al. 2007).

One such approach to designing feedback intervention messages has been motivationally designed feedback messages (Narciss, 2008). We can contextualize this within the theories of self-regulated learning, where the primary function of this feedback rests in guiding the learner to successfully regulate his or her learning process (Butler & Winne, 1995; Narciss, 2008). However, another consideration for this particular study was the target population: USMA cadets. When designing motivational feedback messages for a military population, the theoretical constructs used to ground these designs should reflect the unique profile of this particular group. Therefore, after a review of the literature, the strategies of designing motivational feedback for a military population included the following three theoretical constructs: motivation based on control-value, social identity, and self-efficacy. Experiment 2, then, examined what motivational feedback condition yielded statistically significant positive learning gains. From this experiment, the best condition (self-efficacy messages) were subsequently used in the culminating Experiment 3, which compared multiple detectors for triggering messages. This experimental series is summarized in the below table, each of which is described in the coming sections (see Table 1).

Experiment 1: Multimodal Affect Detection Corpus Collection

Overview

The first experiment of this project took place in September 2013 and was conducted with a group of first-year cadets at USMA to investigate the relationship between affect, behavior, and learning within a modified version of the Tactical Combat Casualty Care

Table 1 Overview of three studies between September 2013–March 2016

Studies	Demographics	Conditions
Experiment 1: Fall 2013, Baseline and Detector Development	Male: n = 99 Female: n = 20	One condition for all participants
Experiment 2: Fall 2015, Motivational Feedback Experiment	Male: n = 110 Female: n = 14	Condition 1, Control-Value Motivational Messages: n = 26 Condition 2, Social Identity Motivational Messages: n = 26 Condition 3, Self-Efficacy Motivational Messages: n = 24 Condition 4, Non-Motivational Feedback Messages: n = 25 Condition 5, No Messages: n = 23
Experiment 3: Spring 2016, Integrated Detectors and Interventions	Male: n = 77 Female: n = 13	Condition 1, Interaction-Based, Sensor-free Detectors: n = 31 Condition 2, Kinect-Based Sensor Detectors: n = 31 Condition 3, Fixed Schedule Feedback Control: n = 28

(TC3Sim) training course developed by the US Army (Sotomayor 2010), as delivered by GIFT. The experiment was conducted in a sensor-rich classroom environment: each cadet was assigned to a research station that consisted of a laptop and several affect sensors. Throughout the study, field observers made regular observations of participant affect and behavior by following a quantitative field observation protocol, BROMP (Ocumpaugh et al. 2012, 2015a). The study yielded results identifying a negative correlation between frustration and student learning outcomes (among a set of affective states studied). Further, this study provided the baseline data needed to create the sensor-based and interaction-based affect detectors to address frustration while participants used TC3Sim. The resulting interaction-based and sensor-based frustration detectors were utilized in subsequent experiments conducted in the following years of the project.

Sensor-Rich Classroom Setup

During the first study, ten separate research stations were configured to collect data simultaneously; each station was used by one cadet at a time. Each station consisted of an Alienware laptop, a Microsoft Kinect for Windows v1.0 sensor, and an Affectiva Q-Sensor, as well as a mouse and pair of headphones.

Kinect sensors recorded participants' physical behavior during the study, including head movements and posture shifts. Each Kinect sensor was mounted on a tripod and positioned in front of a participant. For detailed information on Kinect interaction and algorithms please refer to Paquette et al. (2015).

Q-Sensors recorded participants' physiological responses to events during the study. The Q-Sensor is a wearable arm bracelet that measures participants' electrodermal activity (i.e., skin conductance), skin temperature, and its orientation through a built-in 3-axis accelerometer. However, Q-Sensor logs terminated prematurely for a large number of participants due to a logging issue. For this reason, we omit Q-Sensor data from the remainder of this work.

Sample

There were 119 cadets from USMA who participated in the study (83% male, 17% female). The participants were enrolled in either USMA's PL100: General Psychology, PL 150: Advanced General Psychology, or PL300: Military Psychology courses and recruited through West Point's SONA System. The age of the cadets ranged between 18 and 22.

Method

All participants completed the same training module. The tasks in this experiment were related to standard, simulated combat medic training. The training materials provided to the participants pertained to the knowledge and procedures of hemorrhage control during care under fire and tactical field care, which are all components of TC3Sim (Parsons and Mott 2005). The specific content that participants interacted with was edited content from the TC3Sim training program which consisted of several scenarios. The first scenario was an easy-to-solve scenario about leg amputation and tourniquet

usage. The second scenario was a more complex village scenario with added elements of enemy fire and loud explosions, including a straight-forward leg amputation task and a chest bullet-wound task. The last scenario was newly devised for the set of studies, and was dubbed the *Kobayashi Maru.* The *Kobayashi Maru* scenario involved a fallen soldier with multiple hemorrhages who required immediate medical attention. Irrespective of the actions taken by the participant, the soldier expired quickly.

Additional study materials consisted of pre-tests, training materials, and post-tests, all of which were administered through GIFT. At the onset of each study session, learners completed a content pre-test on tactical combat casualty care. Afterward, participants were presented with a PowerPoint presentation about tactical combat casualty care (see Fig. 2). After completing the PowerPoint, participants completed a series of training scenarios in the TC3Sim serious game where they applied skills, procedures, and knowledge presented in the PowerPoint. In this study, a standard set of scenarios was used (a custom set was used in the following study).

In TC3Sim, the learner adopts the role of a combat medic faced with a situation where one (or several) of her fellow soldiers has been seriously injured. The learner is responsible for properly treating and evacuating the casualty, while following appropriate battlefield doctrine. After the TC3Sim training scenarios, participants completed a post-test, which included the same series of content assessment items as the pre-test. This design decision was made due to the lack of feedback on performance on the pre-test, and the challenge of identifying items of identical difficulty in order to counter-balance two exams. In addition, participants completed two questionnaires about their experiences in TC3Sim: the Intrinsic Motivation Inventory (IMI) (Ryan 1982) and Presence Questionnaire (Witmer and Singer 1998; Witmer et al. 2005). All combined study activities lasted approximately one hour; participants typically spent 17–25 min within the TC3Sim training scenarios.

In addition to pre-tests, post-tests and survey data, interaction data was collected while students used the TC3Sim learning environment. This data contains a list of all actions executed by the student when completing scenarios in TC3Sim, such as checking the casualty's vitals, applying a tourniquet or requesting evacuation, as well as relevant events that occurred during the scenario, such as changes in the casualty's

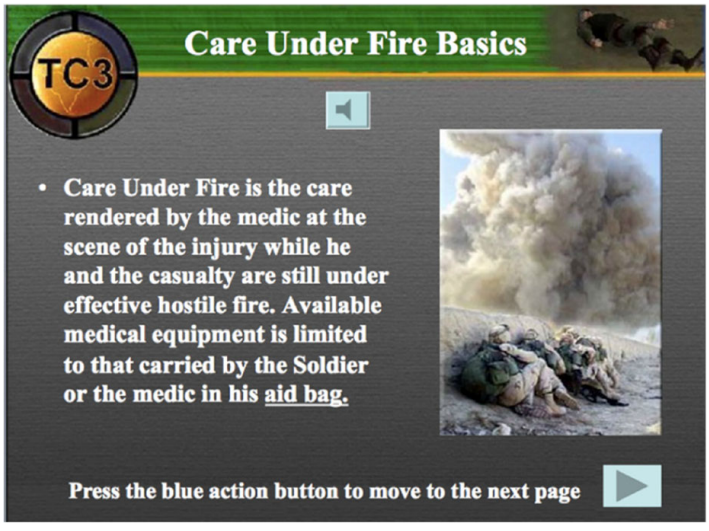

Fig. 2 Screenshot of TC3Sim course PowerPoint material

blood volume and heart rate. In this data, each event and action also have a timestamp associated to indicate the precise moment at which each action or event occurred.

Quantitative Field Observation Protocol

We obtained ground-truth labels of affect using quantitative field observations collected using the Baker-Rodrigo-Ocumpaugh Monitoring Protocol (BROMP) (Ocumpaugh et al. 2012, 2015a).

Affect Detector Model Construction

The BROMP observations collected while cadets were using TC3Sim were used to develop machine-learned models to automatically detect the cadet's affective states. In this section, we discuss our work to develop affect detectors based on cadets' interactions logs and sensor traces.

In total, 3066 BROMP observations were collected by two coders. Those observations were collected over the full length of the cadets' participation in the study, including when they were completing the pre and post-tests, reviewing the PowerPoint presentation, and using TC3Sim. For this study, we used only the 755 observations that were collected while cadets were using TC3Sim. Of those 755 observations, 735 (97.4%) were coded as the cadet being on-task (M across participants = .974, SD = .092, Max = 1), 19 (2.5%) as off-task (M across participants = 0.025, SD = .092, Max = 1), 1 (0.1%) as Without Thinking Fastidiously (M across participants = .001, SD = .006, Max = .05), and 0 as intentional friendly fire. 99.4% of participants were on-task at least once, 35.1% of participants were off-task at least once, and 1.3% of participants were acting Without Thinking Fastidiously at least once. Similarly, 435 (57.6%) of the affect labels were coded as concentrating (M across participants = .576, SD = .239, Max = 1), 174 (23.1%) as confused (M across participants = .231, SD = .185, Max = 1), 73 (9.7%) as bored (M across participants = .097, SD = .161, Max = 1), 32 (4.2%) as frustrated (M across participants = .042, SD = .182, Max = 1), 29 (3.8%) as surprised (M across participants = .038, SD = .045, Max = .227) and 12 (1.6%) as anxious (M across participants = .016, SD = .089, Max = 1). 96.8% of participants demonstrated concentration at least once, 77.3% of participants demonstrated confusion at least once, 64.3% of participants demonstrated boredom at least once, 33.1% of participants demonstrated frustration at least once, 24.0% of participants demonstrated surprise at least once, and 16.9% of participants demonstrated anxiety at least once.

Interaction-Based Affect Detectors

Trainee actions within the software were synchronized to BROMP field observations collected using the HART application (Ocumpaugh et al. 2015b) to generate training data for interaction-based affect detectors.

Thirty-nine features of learner interaction with TC3Sim were distilled, including representations of the player's state such as whether the player was taking cover, whether the player was under fire, and whether the player was with their unit or had separated from them. Representations of the patient's state (which degrades over time

as no action or incorrect action is taken) were also included, such as the patient's change in systolic blood pressure and heart rate, whether the patient had an exposed wound, the patient's lung volume, and the patient's remaining blood volume and rate of bleeding. Features were distilled for player actions such as whether the player was checking the patient's vitals, whether the player applied a bandage or tourniquet, whether the player checked the patient's breathing, whether the player conducted a blood sweep, whether the player communicated with the patient, whether the player requested a security sweep, whether the player requested a case evacuation, and whether the player fired their weapon. Finally, we distilled features of time such as the time since the last player action or within-game event.

Detectors were built separately for each affective state. Each detector was validated using 10-fold participant-level cross-validation. In this process, the trainees are randomly separated into 10 groups of equal size and a detector is built using data for each combination of 9 of the 10 groups before being tested on the 10th group. By cross-validating at this level, we increased confidence that detectors will be accurate for new trainees. Oversampling (through cloning of minority class observations) was used to make the class frequency more balanced during detector development. However, performance calculations were made with reference to the original dataset.

Detectors were fit in RapidMiner 5.3 (Mierswa et al. 2006) using six machine learning algorithms that have been successful for building similar detectors in the past (Baker et al. 2012; Pardos et al. 2014): J48, JRip, NaiveBayes, Step Regression, Logistic Regression and KStar. The detector with the best performance was selected for each affective state. Detector performance was evaluated using two metrics: Cohen's Kappa (Cohen 1960) and A' computed as the Wilcoxon statistic (Hanley and McNeil 1982). Cohen's Kappa assesses the degree to which the detector is better than chance at identifying the modeled construct. A Kappa of 0 indicates that the detector performs at chance, and a Kappa of 1 indicates that the detector performs perfectly. A' is the probability that the algorithm will correctly identify whether an observation is a positive or a negative example of the construct (e.g., is the learner bored or not?). A' is equivalent to the area under the ROC curve in signal detection theory (Hanley and McNeil 1982). A model with an A' of 0.5 performs at chance, and a model with an A' of 1.0 performs perfectly. A' was computed at the observation level.

When fitting models, feature selection was performed using forward selection on the Kappa metric. Performance was evaluated by repeating the feature selection process on each fold of the trainee-level cross-validation in order to evaluate how well models created using this feature selection procedure perform on new and unseen test data. The final models were obtained by applying the feature selection to the complete dataset.

Posture-Based Affect Detectors

The second set of affect detectors were built using data on cadet posture during training with TC3Sim. Kinect sensors produced data streams that were utilized to determine learner posture. Using machine learning algorithms, models were trained to recognize affective states based on postural features.

GIFT's *Sensor Module* was responsible for managing all connected sensors and associated data streams. This includes Kinect sensor data, which is comprised of four complementary data streams: face tracking, skeleton tracking, RGB channel, and depth

channel data. Face- and skeleton-tracking data were written to disk in CSV format, with rows denoting time-stamped observations and columns denoting vertex coordinates. RGB and depth channel data were written to disk as compressed binary data files. To analyze data from the RGB and depth channels, one must utilize the GiftKinectDecoder, a standalone utility that is packaged with GIFT, to decompress and render the image data into a series of images with timestamp-based file names. Data from all four channels can be accessed and analyzed outside of GIFT. For the present study, we utilized only vertex data to analyze participants' posture. Each observation in the vertex data consisted of a timestamp and a set of 3D coordinates for 91 vertices, each tracking a key point on the learner's face (aka face tracking) or upper body (aka skeletal tracking). The Kinect sensor sampled learners' body position at a frequency of 10–12 Hz.

It was necessary to clean the Kinect sensor data in order to remove anomalies from the face and skeletal tracking. Close examination of the Kinect data revealed periodic, and sudden, jumps in the coordinates of posture-related vertices across frames. These jumps were much larger than typically observed across successive frames, and they occurred due to an issue with the way GIFT logged tracked skeletons: recording the most recently detected skeleton, rather than the nearest detected skeleton. This approach to logging skeleton data caused GIFT to occasionally log bystanders standing in the Kinect's field of view rather than the learner using TC3Sim. In these studies, such a situation could occur when a field observer walked behind the trainee.

In order to identify observations that corresponded to field observers rather than participants, Euclidean distances between subsequent observations of a central vertex were calculated. The distribution of Euclidean distances was plotted to inspect the distribution of between-frame movements of the vertex. If the Kinect tracked field observers, who were physically located several feet behind participants, the distribution was likely to be bimodal. In this case, one cluster would correspond to regular posture shifts of a participant between frames, and the other cluster corresponded to shifts between tracking participants and field observers. This distribution could be used to identify a distance threshold for determining which observations should be thrown out, as they were likely due to tracking field observers rather than participants. Although the filtering process was successful, the need for this process reveals a challenge to the use of BROMP for detectors eventually developed using Kinect or video data.

In addition to cleaning the face and skeleton mesh data, a filtering process was applied to remove data unnecessary for the creation of posture-based affect detectors. A majority of the facial vertices recorded by the Kinect sensor were not necessary for investigating trainees' posture. Of the 91 vertices recorded by the Kinect sensor, only three were utilized for posture analysis: top_skull, head, and center_shoulder. These vertices were selected based on prior work investigating postural indicators of emotion with Kinect data (Grafsgaard et al. 2012).

Finally, BROMP observations were synchronized with the data collected from the Kinect sensor. As was the case for our interaction-based detector, the Kinect data provided by GIFT was synchronized to the same Network Time Protocol time server as the BROMP data. This allowed the association of field observations with observations of face and skeleton data produced by the Kinect sensor.

We used the Kinect face and skeleton vertex data to compute a set of predictor features for each field observation. The engineered features were inspired by related

work on posture sensors in the literature on affective learning environments, including work with pressure-sensitive chairs (Cooper et al. 2009; D'Mello and Graesser 2009) and, more recently, Kinect sensors (Grafsgaard et al. 2012).

We computed a set of 73 posture-related features. The feature set was designed to emulate the posture-related features that had previously been utilized in the aforementioned posture-based affect detection work (D'Mello and Graesser 2009; Grafsgaard et al. 2014). For each of three retained skeletal vertices tracked by the Kinect (head, center_shoulder, and top_skull), we calculated 18 features based on multiple time window durations. These features are analogous to those described in Grafsgaard et al. (2012), and were previously found to predict learners' retrospective self-reports of frustration and engagement.

Specifically, we calculated the following features for each vertex:

- Most recently observed distance (i.e., Euclidean distance of vertex from Kinect sensor)
- Most recently observed depth (i.e., Z coordinate of vertex only)
- Minimum observed distance observed thus far
- Maximum observed distance observed thus far
- Median observed distance observed thus far
- Variance in distance observed thus far

We also calculated the minimum distance, maximum distance, median distance, and variance in distance observed during the past 5 s, 10 s, and 20 s. Next, we induced several *net_change* features, which are analogous to those reported by D'Mello and Graesser (2009), as well as Cooper et al. (2009), using pressure-sensitive seat data:

$$net\_dist\_change[t] = \left| \begin{array}{c} head\_dist[t]-head\_dist[t-1]+ \\ cen\_shldr\_dist[t]-cen\_shldr\_dist[t-1]+ \\ top\_skull\_dist[t]-top\_skull\_dist[t-1] \end{array} \right| \quad (1)$$

$$net\_pos\_change[t] = \left| \begin{array}{c} head\_pos[t]-head\_pos[t-1]+ \\ cen\_shldr\_pos[t]-cen\_shldr\_pos[t-1]+ \\ top\_skull\_pos[t]-top\_skull\_pos[t-1] \end{array} \right| \quad (2)$$

These features were calculated from Kinect vertex tracking data, as opposed to seat pressure data. Specifically, the net_dist_change feature was calculated as each vertex's net change in distance (from the Kinect sensor) over a given time window, and then summed together. The net_pos_change feature was calculated as the Euclidean distance between each vertex's change in position over a given time window, and then summed together. Both the net_dist_change feature and net_pos_change feature were calculated for 3 s and 20 s time windows.

We also calculated several sit_forward, sit_back, and sit_mid features analogous to Cooper et al. (2009), and Grafsgaard et al. (2012). To compute these features, we first calculated the average median distance of participants' head vertex from each Kinect sensor. This provided a median distance for each of the 10 study stations. We also

calculated the average standard deviation of head distance from each sensor. Then, based on the station-specific medians and standard deviations, we calculated the following features for each participant:

$$sit\_forward = \begin{cases} 1 & if \ \ head\_dist \leq median\_dist - st\_dev \\ 0 & otherwise \end{cases} \quad (3)$$

$$sit\_back = \begin{cases} 1 & if \ \ head\_dist \geq median\_dist + st\_dev \\ 0 & otherwise \end{cases} \quad (4)$$

The sit_mid feature was the logical complement of sit_forward and sit_back; if a learner was neither sitting forward, nor sitting back, they were considered to be in the sit_mid state. We also computed predictor features that characterized the proportion of observations in which the learner was in a sit_forward, sit_back, or sit_mid state over a window of time. Specifically, we calculated these features for 5, 10, and 20 s time windows, as well as over the entire session to-date.

Posture-based detectors of affect were built using a process analogous to the one used to build our interaction-based detectors. As such, separate detectors were, once again, built for each individual affective state and behavioral construct. All observations labeled as '?' were removed from the training set as they represent observations where the cadet's affective state could not be determined.

Each detector was validated using 10-fold participant-level cross-validation. Oversampling was used to balance class frequency by cloning minority class instances, as was the case when training our interaction-based detectors. RapidMiner 5.3 was used to train the detectors using several different classification algorithms: J48 decision trees, naïve Bayes, support vector machines, logistic regression, and JRip. When fitting posture-based affect detection models, feature selection was, once again, performed through forward selection using a process analogous to the one used for our interaction-based detectors.

Affect Detector Results

As discussed above, each of the interaction-based and posture-based detectors of affect were cross-validated at the participant level (10 folds) and performance was evaluated using both Kappa and A'. Results from the best-performing models for each affective state are shown in Tables 2 and 4. Results from all of the classifiers are shown in Tables 3, 4 and 5. Performance of our interaction-based detectors was highly variable across affective states. The interaction-based detector of boredom achieved the highest performance (Kappa = 0.469, A' = 0.848) while some of the other detectors achieved very low performance. This was the case for the confusion detector that performed barely above chance level (Kappa = 0.056, A' = 0.552). Detectors of frustration and surprise achieved relatively low Kappa (0.105 and 0.081 respectively), but good A' (0.692 and 0.698 respectively). Performance for engaged concentration achieved a Kappa closer to the average (0.156), but below average A' (0.590).

Table 2 Performance of best interaction-based detectors of affect

Affect	Classifier	Kappa	A'
Boredom	Logistic Regression	0.469	0.848
Confusion	Naïve Bayes	0.056	0.552
Engaged Concentration	Step Regression	0.156	0.590
Frustration	Logistic Regression	0.105	0.692
Surprise	KStar	0.081	0.698

In general, posture-based detectors performed only slightly better than chance, with the exception of the surprise detector, which actually performed worse than chance. The boredom detector, induced as a logistic regression model, achieved the highest predictive performance (Kappa = 0.109, A' = 0.528), induced as a logistic regression model. The frustration detector, built as a support vector machine, achieved Kappa = 0.061, A' = 0.518.

Following these analyses, Pearson correlations were computed between the frequencies of each affective state and learning. Pearson correlations were used after first verifying that the data was approximately normally distributed, with skewness and kurtosis well under maximum limits. In general, there was a limited degree of learning (mean = −0.04). The relationships between learning outcomes and confusion and surprise and engaged concentration were not observed to be statistically significant (confusion, $r = -0.107$, $n = 100$, $p = 0.286$; surprise, $r = -0.134$, $n = 100$, $p = 0.180$; engaged concentration, $r = 0.12$, $n = 100$, $p = 0.247$). However, frustration was marginally significant and negatively correlated with learning outcomes (frustration, $r = -0.169$, $n = 100$, $p = 0.092$). Unexpectedly, boredom was positively correlated with learning outcomes (boredom, $r = 0.200$, $n = 100$, $p = 0.046$). Given the negative correlation between our measures of frustration and learning gains, we decided to study the effect of providing feedback to frustrated students in order to help them persevere through the learning activity and mitigate the effects of frustration on learning gains.

Table 3 Performance of all interaction-based detectors of affect

	Eng. Conc.	Boredom	Frustration	Confusion	Surprise
J48	Kappa = 0.077 A' = 0.517	Kappa = 0.312 A' = 0.669	Kappa = 0.006 A' = 0.504	Kappa = 0.053 A' = 0.539	Kappa = −0.015 A' = 0.491
JRip	Kappa = 0.123 A' = 0.539	Kappa = 0.273 A' = 0.659	Kappa = −0.010 A' = 0.494	Kappa = 0.030 A' = 0.512	Kappa = 0.070 A' = 0.543
NaïveBayes	Kappa = 0.170 A' = 0.542	Kappa = 0.434 A' = 0.819	Kappa = 0.099 A' = 0.709	Kappa = 0.056 A' = 0.552	Kappa = 0.022 A' = 0.527
Step Regression	Kappa = 0.156 A' = 0.590	Kappa = 0.409 A' = 0.804	Kappa = 0.068 A' = 0.637	Kappa = −0.011 A' = 0.514	Kappa = 0.004 A' = 0.515
Logistic Regression	Kappa = 0.105 A' = 0.567	Kappa = 0.469 A' = 0.848	Kappa = 0.105 A' = 0.692	Kappa = 0.031 A' = 0.479	Kappa = −0.025 A' = 0.532
KStar	Kappa = 0.119 A' = 0.546	Kappa = 0.395 A' = 0.704	Kappa = 0.058 A' = 0.661	Kappa = 0.050 A' = 0.550	Kappa = 0.081 A' = 0.698

Table 4 Performance of posture-based detectors of affect

Affect	Classifier	Kappa	A'
Boredom	Logistic Regression	0.109	0.528
Confusion	JRip	0.062	0.535
Engaged Concentration	J48	0.087	0.532
Frustration	Support Vec. Machine	0.061	0.518
Surprise	Logistic Regression	−0.001	0.493

The interaction based detector of frustration, a logistic regression model, was based on an equation using the following weight for each variable (with all variables in reference to a 20 s time window, equivalent to the BROMP observation period):

$$w[\text{Sum } of\ Systolic\ Change] = 1.08$$
$$w[\text{Max } of\ Heart\ Rate] = 1.04$$
$$w[Total\ Number\ of\ Blood\ Sweep\ Actions] = -0.07$$
$$w[Total\ Number\ of\ Times\ Player\ is\ Out\ of\ Cover] = 0.131$$
$$w[Total\ Number\ of\ Times\ Player\ is\ Safe\ From\ (Ongoing)\ Enemy\ Fire] = 0.013$$

As the weights show, detection of player frustration was a function of whether the player was being affected by enemy action (with the player being in danger more associated with frustration than simply having active enemies), the status of the patient (with dangerously high patient heart rate and worsening in patient status associated with frustration), and the player's actions to diagnose the patient (associated with less frustration).

Discussion of Initial Affect Detector Modeling Results

Across affective states, the posture-based detectors achieved lower predictive performance than the interaction-based detectors. In fact, the posture-based detectors performed only slightly better than chance, and in the case of some algorithms and

Table 5 Performance of all posture-based detectors of affect

	Eng. Conc.	Boredom	Frustration	Confusion	Surprise
J48	Kappa = 0.087 A' = 0.532	Kappa = −0.019 A' = 0.489	Kappa = −0.010 A' = 0.494	Kappa = −0.010 A' = 0.500	Kappa = −0.030 A' = 0.480
JRip	Kappa = 0.023 A' = 0.518	Kappa = 0.033 A' = 0.503	Kappa = 0.001 A' = 0.494	Kappa = 0.062 A' = 0.535	Kappa = −0.005 A' = 0.495
Logistic Regression	Kappa = −0.003 A' = 0.490	Kappa = 0.109 A' = 0.528	Kappa = 0.050 A' = 0.512	Kappa = 0.011 A' = 0.502	Kappa = −0.001 A' = 0.493
Naïve Bayes	Kappa = 0.005 A' = 0.498	Kappa = 0.035 A' = 0.522	Kappa = 0.006 A' = 0.503	Kappa = 0.017 A' = 0.507	Kappa = −0.024 A' = 0.490
SVM	Kappa = −0.026 A' = 0.327	Kappa = 0.095 A' = 0.534	Kappa = 0.061 A' = 0.518	Kappa = 0.034 A' = 0.462	Kappa = −0.008 A' = 0.492

emotions, worse than chance. This finding is notable, given that our distilled posture features were inspired largely from the research literature, where these types of features have been shown to predict learner emotions effectively in other contexts (Cooper et al. 2009; D'Mello and Graesser 2009; Grafsgaard et al. 2014).

There are several possible explanations for why the posture-based predictors were not more effective. First, our use of BROMP to generate affect labels distinguishes our work from prior efforts, which used self-reports (Cooper et al. 2009; Grafsgaard et al. 2012, 2014) or retrospective video freeze-frame analyses (D'Mello and Graesser 2009). It is possible that BROMP-based labels of affect present distinct challenges for posture-based affect detection. BROMP labels are based on holistic judgments of affect, and pertain to 20-s intervals of time, which may be ill matched for methods that depend upon low-level postural features to predict emotion. Similarly, much of the work on posture-based affect detection has taken place in laboratory settings involving a single participant at a time (D'Mello and Graesser 2009), especially prior work using Kinect sensors (Grafsgaard et al. 2012, 2014). In contrast, our study was performed with up to 10 simultaneous participants, introducing potential variations in sensor positions and orientations. This variation may have introduced noise to our posture data, making the task of inducing population-general affect detectors more challenging than in settings where data is collected from a single sensor. If correct, this explanation underscores the challenges inherent in scaling and generalizing sensor-based affect detectors. One direction for addressing these sources of variation in the future is to investigate alternate positioning for the Kinect sensors, including use of side angles (Sanghvi et al. 2011). Additional directions include engineering predictor features using standard units rather than Cartesian space, as well features that account for individual differences in body shape and size (Worsley et al. 2015).

Another possible explanation for the limited success of sensor-based detectors is that there were population-specific norms of restraining non-verbal displays of emotion. Both BROMP observers noted that the USMA cadets' affective expressiveness was generally different in kind and magnitude than the K-12 and civilian academic populations they were more accustomed to studying. Specifically, they indicated that the USMA population's facial and behavioral expressions of affect were relatively subdued, perhaps due to military cultural norms. While cadets did display noticeable shifts in posture during the study, displays of affect via movement and body language may have been more difficult to recognize by this type of sensor than would have otherwise been the case in other populations.

In general, we consider the study population, BROMP affect labels, and naturalistic research setup to be strengths of the study. Indeed, despite the difference in how military trainees display affect compared to the K-12 and civilian academic population, human observers were able to achieve the inter-rater reliability required by BROMP (Kappa > = 0.6) (Ocumpaugh et al. 2012, 2015a). Thus we do not have plans to change these components in future work. Instead, it may be worth making efforts to revise and enhance the data mining techniques that we employ to recognize learner affect, as well as the predictor features engineered from raw posture data.

It is notable that our interaction-based detectors had a more varied performance than had been seen in prior studies using this methodology; the detectors were excellent for boredom, and varied from good to just above chance for other constructs. It is possible that this too is due to the population studied, but may also be due to the nature of the features that were distilled in order to build the models.

It is also possible that some of the affective states for which interaction-based detection was less effective and simply did not manifest consistently in the interactions with the learning environment across different trainees. It is thus difficult to determine whether poor performance of detectors was due to insufficient feature engineering or inconsistent behaviors by the trainee. As such, the creation of interaction-based detectors is an iterative process, where features are engineered, and models are induced and refined, until performance reaches an acceptable level, or no improvement in performance is observed, despite repeated knowledge-engineering efforts.

We aim to identify methods to improve the predictive accuracy of interaction-based and posture-based detectors in future work. Notably, one advantage that posture-based detectors possess relative to interaction-based detectors is that posture-based detectors may be more generalizable, since they pertain to aspects of learner behavior that are outside of the software itself. By contrast, much of the effort invested in the creation of interaction-based detectors is specific to the system for which the detectors are created. Features are built to summarize the learner's interaction in the learning environment and, as such, are dependent on the system's user interface. Much of the creation of interaction-based detectors must hence be replicated for new learning environments, though there have been some attempts to build toolkits that can replicate features seen across many environments, such as unitizing the time between actions by the type of action or problem step (e.g., Rodrigo et al. 2012).

On the other hand, the process of creating predictor features for posture-based detectors requires considerable effort when compared with building a set of features for interaction-based detectors, such as elaborate efforts to adequately clean the data, but at least in principle, it is only necessary to develop the methods for doing so once. The same data cleaning and feature distillation procedures can be repeated for subsequent systems. This is especially useful in the context of a generalized, multi-system tutoring framework such as GIFT. Although different posture-based affect detectors might need to be created for different tutoring systems—due to differences in the postures associated with affect for different populations of learners, environments and contexts—the posture features we computed from the data provided by Kinect sensors will ultimately become available for reuse by any tutor created using GIFT. This has the potential to considerably reduce the time required to build future posture-based affect detectors for learning environments integrated with the GIFT architecture.

A limitation of the study was usage of the same pre-test and post-test for assessing cadet learning gains. The study drew upon a pre-existing bank of assessment questions on tactical combat casualty care, which were limited in number and scope. The question bank was sufficient for generating a single test, but there were not enough distinct items to populate several balanced assessments with distinct questions. A potential future direction for this work is devising an expanded bank of assessment questions, as well as investigate complementary techniques for assessing knowledge gains in the tactical combat casualty care domain (e.g., stealth assessments).

In sum, the first experiment yielded sensor-free, interaction based affect detectors as well as posture sensor-based detectors. With these two models of frustration detectors built, the next phase of the project included devising the feedback messages that that would be delivered to participants while engaged with TC3Sim and triggered by these detectors that were subsequently built into GIFT.

Experiment 2: Testing Frustration Feedback in Gift

The next phase of the multi-study project was realized in a second experiment run in September 2015. In this experiment, the interaction based frustration detectors were integrated into GIFT in order to respond to participants' frustration while engaged in TC3Sim. Three distinct motivational feedback interventions were tested in the experiment to determine which feedback intervention messages yielded the best learning gains.

Addressing Frustration in ITSs

In a survey of the literature, it was noted that when a learner is in a frustrated state in ITSs there are a range of solutions that can be employed to address this affective state. These solutions include changing the elements in a system that elicit frustration, as well as supporting the learner in their ability to recover, manage, and persist in their task (Klein et al. 2002; Kapoor et al. 2007).

Amsel's (1992) frustration theory supports the notion that goal attainment includes overcoming emotional conflict rather than avoiding emotional conflict. This theory supports the idea of forgoing explicit feedback interventions that would quickly relieve the frustration of the participant, and instead providing motivational feedback that would support participants' efforts to persist through frustration (de Wit and Dickinson 2015; Marien et al. 2015). As such, when considering developing motivational feedback messages to be used as an intervention upon the detection of frustration, it was determined that messages should be meaningful, relevant, and support goal-oriented efforts. Therefore, the intervention feedback messages were designed according to three theories of motivation. These theories were: control-value messages (Pekrun et al. 2006) which would indicate the value of the activity, i.e., playing TC3Sim to learn combat medic skills; social identity messages (Tajfel and Turner 1979) which would target the military identity of the participants; and self-efficacy messages (Bandura 1986) which would support the participants' ability to achieve a learning goal through effort.

A Theory-Based Approach to Reducing Student Frustration

The messages designed for this second experiment were devised and/or adopted because they were linguistically congruent with one of the three motivational theories adopted for the feedback interventions used to address participant frustration. Each of the three interventions was also selected with the goal of being fail-soft, i.e. having minimal negative impact if delivered based on a false positive from the detector. For the full list of feedback messages used in this experiment, please refer to Appendix.

Control-value theory was developed by Pekrun et al. (2006) as a comprehensive, integrative approach to understanding emotions in education. When individuals feel in or out of control of achievement activities and outcomes that are subjectively important to them, they experience specific achievement emotions (Frenzel et al. 2007). Control-value motivational messages, then, were designed around the idea that achievement emotions such as frustration can be influenced by changing the student's subjective perception of control and value within the learning environment. An example of the

kind of message devised for this condition is as follows: "A 2008 study from a hospital in Baghdad found an 87% survival rate with use of tourniquets."

The social identity message design capitalized on Tajfel and Turner's (1979) social identity theory, which states our identities are formed through the groups to which we belong, creating some degree of uniformity of perception and action that exist among group members. Social identity is aligned with the situated social cognition perspective that proposes cognition and action are not discrete entities but dynamically shaped by each other (Schwarz 2007, 2009; Smith and Semin 2004, 2007). The social-identity condition messages, then, capitalized on the notion that the cadets were members of the military, and under the social identity theory, people prefer identity-congruent to identity-incongruent actions. These messages suggest that appropriate responses to frustration are identity-congruent. An example of the kind of message devised for this condition is as follows: "Every single man in this Army plays a vital role," said General Patton. "Don't ever let up. Every man has a job to do and he must do it."

The third theory of motivation used to develop motivational feedback messages involved creating messages framed by the theory of self-efficacy (Bandura 1986). Self-efficacy includes how the learner seems themselves as an individual, and their ability to succeed in a task if they persist (Bandura 1986). The self-efficacy condition messages, subsequently, were designed to persuade the learner that they had the necessary skills to succeed; an example of which is as follows: "Difficult doesn't mean impossible. It means work harder till your combat mission is achieved."

There was also a fourth bank of messages devised for this study that served as one of the two control conditions in the experiment. This fourth bank of messages were non-motivational messages: factoids related to hemorrhage and bleeding control, as well as control and use of tourniquets. An example of this non-motivational message is as follows: "The modern combat medic has its roots in the American Civil War, when enlisted soldiers served as hospital stewards."

Integration and Implementation of Interaction-Based Frustration Detector

In order to provide motivational feedback when frustration was detected during training with TC3Sim, the interaction-based affect detector of frustration was integrated with GIFT. This involved implementing the following mechanisms in GIFT: (1) accumulate interaction data over a period of time; (2) compute a set of features summarizing the interaction data; and (3) apply a classifier model created using RapidMiner to the summary features (Paquette et al. 2015).

First, GIFT accumulates interaction data over a predetermined period of time. This is achieved by having GIFT listen to the events broadcast by the learning environment, identify the events related to student actions, and collect a list of those actions. In this experiment, interaction data from TC3Sim was collected over periods of 20 s—a period of time that is consistent with the observation time in BROMP.

Second, once GIFT has accumulated a 20-s record of student actions, it computes summary features of the student's behavior using a custom-authored Python script. In our experiment, this script takes the list of actions executed by the participant in TC3Sim and computes the value for each of the features that are required by the frustration detection model.

Third, GIFT instantiates a RapidMiner process that uses the set of computed feature as an input. This RapidMiner process is responsible for applying the appropriate classifier model to the set of computed features. In our experiment, a RapidMiner process loads the frustration detector, applies it to the set of summary features, and outputs the detector's confidence of whether the student is currently frustrated or not. This confidence value is interpreted by GIFT in order to determine whether the student is likely to be in a state of high frustration. In this experiment, a confidence threshold of 0.50 was used to distinguish moments of high detected frustration. When frustration was detected, a motivational feedback message was provided.

Overview of Experiment

The second experiment in this project focused on examining whether motivational feedback messages delivered during TC3Sim training, and upon the detection of high frustration, would positively impact the learning outcomes of USMA cadets. What follows is the methodology employed to compare these motivational feedback interventions against two control conditions: the non-motivational message condition, and a condition with no messages delivered at all.

Participants

Log files were extracted from GIFT from all 141 participants, and were examined to ensure all files were complete data sets. 17 participants' log files had a gap in the output where the participant either did not have a pre-test or post-test – a result of failures of the laptop and/or GIFT. Subsequently, these 17 participants were dropped from the data analysis, as it was impossible to calculate learning gains from incomplete sets of data.

In total, the final data analysis was run on 124 participants: 26 participants were in the first condition, 26 participants in the second condition, 24 participants in the third condition, 25 participants were in the first control condition, and 23 participants in the second control condition.

Method

Data collection was conducted over a three-day period at USMA. Experiment 2's methodology was similar to that of Experiment 1, with the following differences: (1) addition of experimental conditions; (2) omission of Kinect and Q-Sensor data collection; (3) two additional scenarios in TC3Sim, which are next described; (4) administration of Short-Grit Scale (Duckworth and Quinn 2009); (5) a longer pre- and post-test using the same 10 questions from the pre and post-test used in Experiment 1, but adding ten additional pre and post test questions. These changes to the pre- and post-tests were made to see whether there would be an impact on the direction of learning gains in the self-efficacy condition, as Slack (2014) had previously demonstrated a positive correlation between the constructs of self-efficacy and the Short Grit Scale (Duckworth and Quinn 2009). And more generally, we made these changes to determine whether there would be an overall interaction effect of grit on condition and learning gains, giving some evidence as to whether the trait of grit functioned as a moderator across motivational feedback conditions. In terms of the intervention, one

feedback message was delivered per scenario for all conditions that were designed to deliver feedback messages, upon the detection of frustration, irrespective of the amount of times the frustration detectors detected frustration during the scenario. Only the control condition of *No Messages* (control condition 2) did not deliver a feedback message upon the detection of frustration. Lastly, while some BROMP data was collected in this experiment, due to equipment failure the observations had to be discontinued multiple times during the three-day experiment, and these measures were not used in the final data analysis.

Equipment and Materials

The content that participants first interacted with was the same edited content from the TC3Sim training program used in Experiment 1. However, some scenarios were repeated multiple times. The first (same as study 1) scenario was an easy-to-solve scenario: a leg amputation, which required the application of a tourniquet. The second (third in study 1) scenario was the *Kobayashi Maru* (a multiple hemorrhage scenario that was devised so that the fallen soldier that required medical attention had multiple wounds and would expire quickly – no matter what actions the participant took). The third (repeated) scenario was a repeat of the leg amputation scenario. The fourth (second in study 1) scenario was a more complex village scenario with added elements of enemy fire and loud explosions, with an easy, winnable leg amputation task, and a chest bullet wound task. The final (repeated) scenario was again the *Kobayashi Maru*, a no-win situation.

It was anticipated that laying out the scenarios in this manner would increase the frequency of frustration among the participants. This increase of frustration would be a result of the fact that while the first scenario would lead participants to believe that this was a game that was winnable, the second no-win *Kobayashi Maru* scenario would undermine those beliefs. Frustration would be further increased in the third scenario by having a repeat of the easy winnable first leg-amputation scenario, followed with a more challenging but again winnable (complex village) fourth scenario. However, the repeat of the final no-win *Kobayashi Maru* scenario would have participants finish the serious game task with an expected increase and more sustained frustrated state, as the participants would have vacillated between confirming their beliefs about winning the game and having those beliefs undermined irrespective of any efforts taken.

Results

The assumptions for normally distributed data for pre- and post-test were met, as were assumptions for one-way analyses of variance (ANOVA) and repeated measures analysis of variance (rANOVA) analyses.

Across all conditions, there was in increase in learning outcomes from pre to post test (see Table 6 and Fig. 3).

The grand mean frequency of detected frustration across all conditions was 6.43 instances of frustration detected while participants engaged with TC3Sim. The condition with the greatest frequency of system-detected frustration was the no message condition (full control condition 2), with a mean frequency of 6.70 instances of detected high frustration. The two conditions with the lowest frequencies detected for high

Table 6 Means chart for pre and post test scores by condition

Conditions	$\bar{x}_{PRE}$	[SD]	$\bar{x}_{POST}$	[SD]	$\bar{x}_{DIFF}$
1_Control Value	13.115	1.966	13.731	2.219	0.61
2_Social Identity	13.885	1.904	14.231	2.286	0.35
3_Self-Effficacy	13.167	2.057	14.333	1.926	1.166
4_1c_NonMotivational	13.56	1.938	14.240	1.715	0.68
5_2c_NoMessages	13.87	1.632	14.304	1.820	0.434

frustration were the control-value condition (condition 1), with a mean of 6.19 detected high frustration events, and the self-efficacy condition (condition 3), with a mean of 6.33 detected high frustration events (see Table 7 and Fig. 4).

The Presence survey from the Intrinsic Motivation Inventory (IMI) (Witmer and Singer 1994) was administered to assess a participants' subjective experience in vMedic. There was an overall (n = 124) mean presence score of 109.573 with a standard deviation of 16.52 (see Table 8 and Fig. 5).

The Short Grit Scale (Duckworth and Quinn 2009) was administered to participants during the experiment. The overall (N = 124) grand mean was 3.80 with a standard deviation of 0.56. The range of means by condition was 3.67 (self-efficacy condition 3) to 3.89 (control-value theory condition 1) (see Table 9 and Fig. 6).

Main Effects Two-way mixed design rANOVA analysis was run to determine if there is a main effect when controlling for frustration and three-way interaction with frustration. There was a significant main effect: (rANOVA): F(4, 114) = 3.68, p = .007, ηp^2 = .114 Then, a two-way mixed design rANOVA analysis was run comparing the motivational conditions (conditions 1, 2, & 3) to the control conditions (conditions 4, & 5). This analysis indicated that motivational conditions had higher learning outcomes

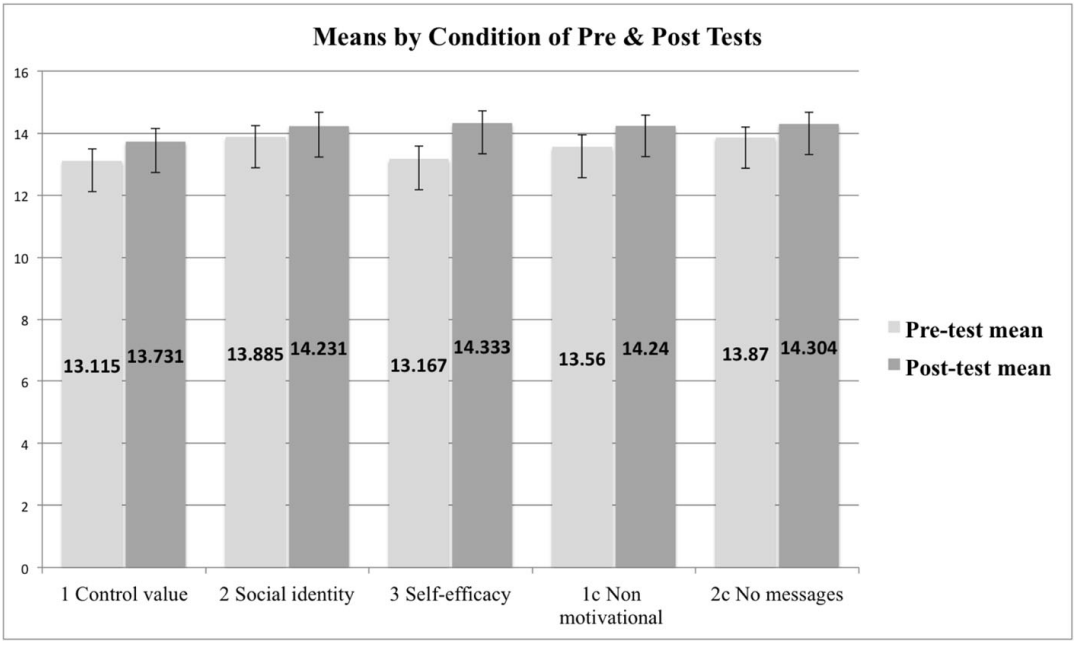

Fig. 3 Means by condition of pre-post tests scores

Table 7 Frequency mean, minimum, and maximum scores of system detected frustration

Condition	N	Mean	Std. Deviation	Min	Max
1_Control-value	26	6.19	1.789	0	8.0
2_Social identity	26	6.65	0.936	4	8.0
3_Self efficacy	24	6.33	1.373	3	8.0
1c_ Non motivational	25	6.28	1.904	0	9.0
2c_No messages	23	6.70	1.222	4	9.0

than the control conditions: (rANOVA): F(1, 120) = 5.627, p = .019. ηp^2 = .045, power = .653. Also, there was a statistically significant interaction between conditions, frustration, and learning outcomes (rANOVA): F(1, 120) = 5.578, p = .020. ηp^2 = .044, power = .649.

Conducting a post-hoc, simple main analysis for a two-way rANOVA with a significant interaction requires running separate rANOVA's for factors under investigation (Keselman 1998; Verma 2016; Weinberg and Abramowitz 2002). Therefore, subsequent independent pairwise rANOVA's were run comparing each motivational condition separately to each control condition. This analysis indicated that the self-efficacy condition (N = 24) had higher learning outcomes than the non-motivational feedback control group (N = 25), (rANOVA): F(1, 45) = 9.945, p = .002, ηp^2 = .181, power = .870, (see Table 10). Using the Benjamini-Hochberg adjusted alpha post-hoc test, these results are still significant: p = .002 < B-H α = .008 (see Table 10).

This analysis also indicated that the self-efficacy condition (N = 24) had higher learning outcomes than the no message control group (N = 23), (rANOVA): F(1, 43) = 7.355, p = .007, ηp^2 = .159, power = .796 (see Table 10). Again, using the Benjamini-Hochberg adjusted alpha post-hoc test, these results are still significant: p = .007 < B-H α = .016 (see Table 10).

No other comparisons were significant when using the Benjamini-Hochberg procedure.

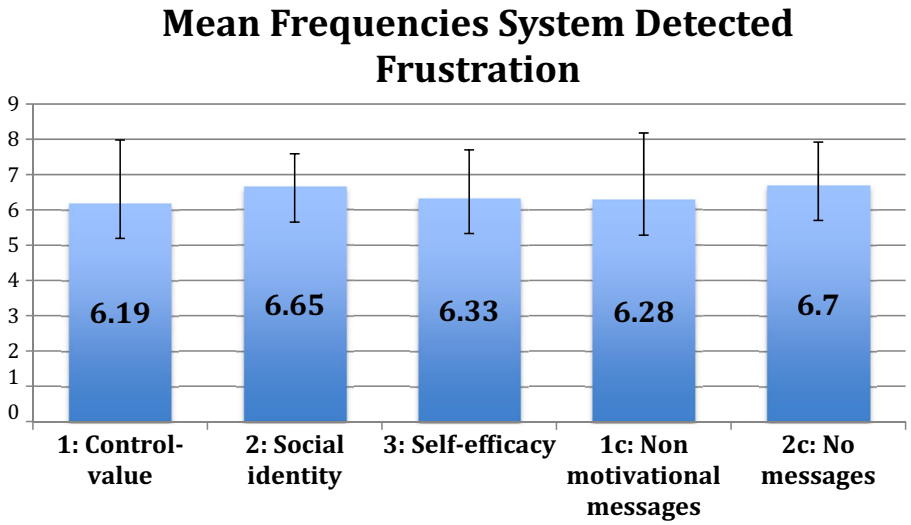

Fig. 4 Mean frequency of system detected frustration by condition

Table 8 Presence mean scores and standard deviation by condition (max possible score: 165)

Condition	N	Mean	Std. Deviation
1_Control-value	26	106.30	17.36
2_Social identity	26	110.27	17.68
3_Self efficacy	24	107.62	17.12
1c_ Non motivational	25	114.28	16.70
2c_No messages	23	109.40	13.26

Presence To test whether presence (measured using the questionnaire discussed above) had a moderating effect between conditions on learning outcomes, a two-way mixed design rANOVA analysis was conducted. Using measures from the Presence questionnaire (Witmer and Singer 1998), presence did not have a statistically significant association with learning outcomes, (rANOVA): F(1114) = 1.639, p = .203, ηp^2 = .014, power = .246, and there was not a statistically significant interaction between presence and condition on pre-post test scores, (rANOVA): F(4114) = 0.162, p = 0.957, ηp^2 = .006, power = .083.

Grit To test whether grit (measured using the questionnaire discussed above) had a moderating effect on learning outcomes across conditions, a two-way mixed design rANOVA analysis was conducted (Table 11). Using measures from the short grit survey (Duckworth and Quinn 2009), there was a statistically significant difference between conditions and positive learning outcomes controlling for grit and interaction of grit by condition and learning outcomes: (rANOVA): F(4114) = 2.631, p = .038, ηp^2 = .085, power = .721. There was also a statistically significant interaction effect of grit by condition and learning outcomes (rANOVA): F(4114) = 2.903, p = .025, ηp^2 = .092, power = .768.

With this significant interaction, an analysis on the simple effects of grit by condition were conducted, running simple main effect analyses separately at each level of condition (Keselman 1998; Verma 2016; Weinberg and Abramowitz 2002). The analysis of the simple means showed that grit had a statistically significant effect on learning

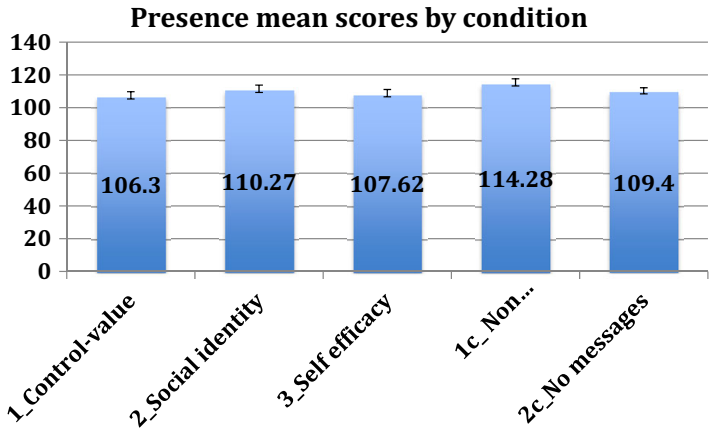

Fig. 5 Presence mean scores and standard error

Table 9 Grit mean scores and standard deviation by condition (max score: 5 = extremely gritty)

Condition	N	Mean	Std. Deviation
1_Control-value	26	3.89	.54
2_Social identity	26	3.84	.49
3_Self efficacy	24	3.67	.70
1c_ Non motivational	25	3.71	.62
2c_No messages	23	3.88	.42

outcomes only within the control-value condition (condition 1), $F(1, 24) = 7.304$, $p = .012$, $\eta p^2 = .233$, power = .737. However, when applying a Benjamini-Hochberg alpha adjustment post-hoc control, this difference can instead be seen as marginal, $p = .012 >$ B-H $\alpha = 0.01$ but $p = .012 <$ B-H $\alpha^*2 = 0.02$ (Table 12).

Yet, in conducting a follow up analysis, splitting the data further into high and low grit groups, using the mean grit value of 3.80, the analysis indicated that low grit participants in the control-value theory condition (condition 1) had higher learning outcomes than high grit participants in the control-value theory condition, (rANOVA): $F(1, 25) = 35.000$, $p = 0.001$, $\eta p^2 = .883$, power = .999. After making Benjamini-Hochberg alpha adjustments, this difference remained significant: $p = .001 <$ B-H $\alpha = 0.005$ (Table 13).

In examining the means of the control-value condition (condition 1) between high and low grit participants, the data revealed that the low grit group had positive learning gains but the high grit group had negative learning gains, although they did not have a statistically significant decline in performance (see Table 14 and Fig. 7).

A closer examination of the pre and post test scores were conducted for the control-value condition, and while there were two participants who scores seem to make them outliers (post test = 9 correct), these participants each had the same score of 9. These participants' data were not removed from the analysis as they were only one point

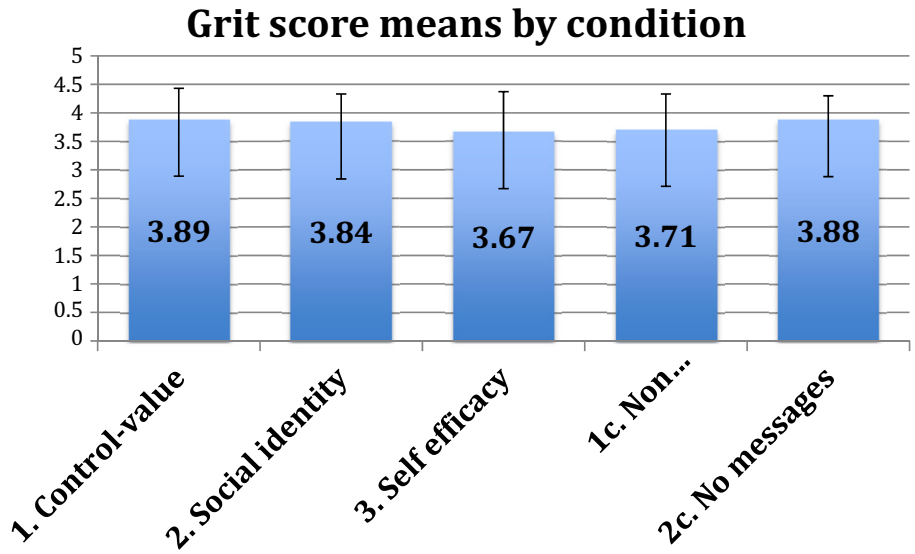

Fig. 6 Grit mean scores by condition

Table 10 Summary of pairwise analyses (rANOVA's) between intervention conditions vs. control groups

Intervention	Control group	df	F	Sig	B-H α	ηp^2	Power
Control-value	Non motivational	1	.004	.948	0.033	.000	.050
Social identity	Non motivational	1	.877	.354	0.041	.018	.151
**Self efficacy*	*Non motivational*	1	9.945	.002	0.008	.181	.870
Control-value	No messages	1	2.290	.137	0.025	.048	.316
Social identity	No messages	1	.352	.556	0.500	.008	.089
**Self-efficacy*	*No messages*	1	7.355	.007	0.016	.146	.755

*statistically significant after Benjamini-Hochberg correction

outside of the normal range of post-test scores: 10–18 correct in the control-value condition.

Specifically, participants with low grit averaged a positive learning gain of +2.495 points, whereas high grit participants had negative learning gains of −0.22 points, (see Table 14). This seems to indicate that the control-value messages had a positive impact on participants with low grit scores, perhaps encouraging them to see the value in the experiment or the learning activity more broadly. For high grit participants, these participants might have seen the messages as unnecessary, annoying, or even frustrating – perhaps even causing some disengagement with the experiment/learning activity.

Discussion of Motivational Feedback Intervention Results

In conclusion, the results of this experiment support previous theories and empirical research that have recognized the need to identify and address affective states that lead to disengagement in learning (Baker et al. 2010; D'Mello and Graesser 2011; D'Mello et al. 2013). These results also contribute to the body of research that has given evidence that providing interventions in the form of feedback messages can positively affect the learning of domain content in ITSs (Wagster et al. 2007; Roll et al. 2011).

Further, this work demonstrates that interaction-based, sensor-free detectors embedded in technology-based learning environments (such as serious video games) can be used to trigger interventions that can lead to better learning outcomes.

Table 11 Means chart rANOVA test with interaction of grit by condition

Conditions	$\bar{x}_{PRE}$	[SD]	$\bar{x}_{POST}$	[SD]	$\bar{x}_{DIFF}$
1 Control Value	13.115	1.966	13.731	2.219	0.61
2 Social Identity	13.885	1.904	14.231	2.286	0.35
3 Self-Efficacy	13.167	2.057	14.333	1.926	1.166
4 1c - NonMotivational	13.56	1.938	14.240	1.715	0.68
5 2c - NoMessages	13.87	1.632	14.304	1.820	0.434

Table 12 Benjamini-Hochberg alpha adjustments pre-post-test scores by condition by grit

*Conditions	*p*-value	B-H α
Control value	0.012	0.01
Social identity	0.686	0.05
Self-efficacy	0.103	0.02
Non-motivational messages	0.679	0.04
No messages	0.119	0.03

*No condition reached significance after adjusting with Benjamini-Hochberg alpha; Control-value condition was marginal

Overall, self-efficacy based motivational feedback interventions were associated with better learning when addressing frustration. While this current study was limited to a military population, it is likely that this motivational approach would work outside of a military population. In particular, over 80% of this study's sample population had not having previously served in the military, rendering the identity of this group arguably closer to undergraduate population of a comparable higher education institution than a population of active military personnel.

A valuable area of future work might be to investigate how this difference occurred. While the difference manifested primarily as differences in learning rather than changes in the prevalence of affect, it might be valuable to study (for example) whether the dynamics of affect change in the presence of these interventions, and whether the changes in affective dynamics may explain the differences in learning outcomes for the self-efficacy condition. Similarly, it may be worth attempting in future work to identify cases where a student received an affective intervention despite not having been frustrated (e.g. a false positive by

Table 13 Benjamini-Hochberg alpha adjustments for significance of pre-post-test scores by condition by high/low grit

Conditions	Grit level	P- value	B-H α
** Control-value*	*Low*	*.001*	*0.005*
Control-value	High	.736	0.04
Social Identity	Low	.232	0.02
Social Identity	High	.865	0.045
Self-efficacy	Low	.376	0.03
Self-efficacy	High	.066	0.01
Non-motivational messages	Low	.235	0.025
Non-motivational messages	High	.403	0.035
No messages	Low	.122	0.015
No messages	High	.916	0.05

**significant after Benjamini-Hochberg alpha adjustments*

Table 14 Means chart rANOVA analysis of high and low grit participants in control-value condition

Conditions	N	Mean Pre	Std. Deviation	Mean Post	Std. Deviation	Sig.
1 Low Grit	8	12.38	1.69	14.88	2.04	0.001
2 High Grit	18	13.44	1.81	13.22	2.24	0.736

the detector), perhaps through triangulating BROMP with the affect detectors, to study the impacts of false positives of affect dynamics and learning.

While presence did not interact with conditions to yield a difference between learning outcomes by condition, there was an interactive effect of grit by condition, providing some marginal evidence that motivational messages were more effective for participants with comparatively low grit measures vs. high grit measures in the control-value condition.

These results contribute towards the body of cognitive performance theory and research by providing empirical evidence for effective approaches to address motivation in simulated learning environments, considerations on the moderating effect of frustration in relation to motivational feedback, as well as evidence regarding tailoring motivational feedback according to trait characteristics such as grit. Specifically, these results include evidence that the self-efficacy motivational feedback messages used to intercede in instances of high frustration can promote greater learning gains than control conditions, and motivational messages based on the theory of control-value may be effective for low grit populations, but may have a negative impact on high grit populations.

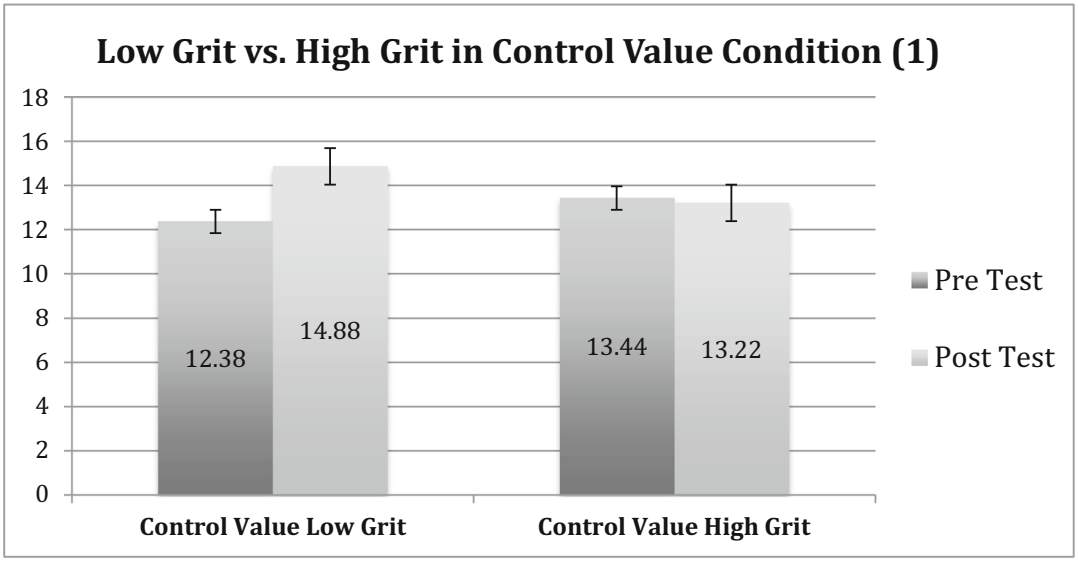

Fig. 7 Pre-post test comparing high /low Grit groups in the control-value condition

Experiment 3: Investigating Affect Detector-Driven Frustration Feedback

Overview

The purpose of the third experiment was to study whether the different methods of detecting frustration for the purpose of delivering self-efficacy feedback intervention messages led to different learning outcomes.

Participants

101 cadets from the USMA participated in the study (85.6% male; 14.4% female). Recruitment was congruent to that of Experiment 1 and 2. The age of the cadets ranged between 18 and 24.

Participants were randomly placed in one of three conditions that all delivered the same self-efficacy messages that were devised and employed in Experiment 2. As in the self-efficacy condition of Study 2, each TC3Sim scenario was assigned a distinct self-efficacy feedback message, and only one self-efficacy feedback message was delivered per scenario. In total, participants spent approximately 15–20 min in the TC3Sim scenarios.

The three conditions of this experiment were as follows:

(1) an interaction-based detector condition that delivered self-efficacy feedback messages upon detection of frustration (n = 31);
(2) a Kinect-based detector condition that delivered self-efficacy feedback messages upon the detection of frustration using the Kinect sensor-based detector (n = 31);
(3) a control condition that delivered the self-efficacy feedback messages at a scheduled time, using the mean delivery times of detected frustration from Experiment 2 (n = 28). (In practice, this resulted in fairly similar delivery times between conditions (1) and (3), as much of the frustration was triggered by specific events in the game, such as the *Kobayashi Maru* scenarios.)

Method

Experiment 3 was held during the first week of March 2016, and it ran for four days. Experiment 3's design and methodology was similar to that of Experiments 1 and 2. Aside from the differences in conditions, the only difference between Experiment 1 and Experiment 3 was the use of shorter pre-tests and post-tests in Experiment 1. Experiment 3 used the same pre- and post-test items (10 items for each instrument), but another ten items were added in order to increase the power of the two assessment instruments. Experiment 3 and Experiment 2 also differ in terms of the test length. Due to logistical limitations, Experiment 3 had the same three scenarios as Experiment 1, rather than the larger number of scenarios seen in Experiment 2. Also, while this study used the same interaction-based, sensor free detectors as were employed in Experiment 2, this study added the integration of posture-based frustration detectors.

Integration and Implementation of Posture-Based Frustration Detector

To enable GIFT to utilize posture data to infer changes in student frustration during TC3Sim training, a posture-based frustration detector that utilized Kinect data was integrated with GIFT. The posture-based detector was implemented using RapidMiner Studio 7.0. The detector was integrated with GIFT following an analogous procedure to the one used for the interaction-based frustration detector.

Model induction involved training the model using the entire dataset collected from Experiment 1—no cross validation was applied—and utilizing the same set of predictor features as described in the previous section on Affect Detector Model Construction. The posture-based detector consisted of a support vector machine (SVM) trained with Weka's implementation of sequential minimal optimization (W-SMO), and it provided binary classifications about the presence (or absence) of student frustration during training with TC3Sim.

In order to integrate posture-based frustration detection through GIFT's Sensor Module, GIFT was configured to listen for, and accumulate, raw Kinect vertex position data; no RGB or depth channel data from the Kinect was utilized for run-time frustration detection, nor was any other type of sensor data. Similar to the interaction-based detector, the posture-based detector involved accumulating a series of vectors of Kinect vertex data over a 20 s window. This included XYZ-coordinate data for all vertices that the Kinect tracked.

After 20 s of Kinect data had been accumulated, the data was provided to a Python script that filtered and transformed the top_skull, head, and center_shoulder vertex data into a vector of predictor features consistent with the format expected by the SVM frustration classifier. The transformed data was returned to GIFT and passed to the run-time RapidMiner process that applied the SVM frustration classifier to the posture feature vector. It should be noted that the posture-based frustration detector was implemented separately from the interaction-based frustration detector; the two models did not work concurrently, nor in combination. GIFT does support run-time multi-channel affect detection, but that was not part of the design for Experiment 3.

Challenges in use of Automated Detectors

An initial review of the GIFT log files of the frequency of system-detected frustration for the interaction-based detector condition (condition 1) and Kinect-based detector condition (condition 2) revealed that the interaction-based detectors intervened on every student and delivered a self-efficacy message in the second and third scenarios of TC3Sim for almost every student, although not the first scenario. It should be noted that the control condition of fixed-schedule delivery (condition 3) was configured to deliver self-efficacy messages at approximately the same time as the interaction-based detector, using the mean time of the delivery from experiment time to inform the time of message delivery.

However, upon close examination of the log files of the Kinect-based condition in Experiment 3, it was determined that the Kinect-based detectors did not trigger the delivery of any self-efficacy messages in any of the three scenarios of TC3Sim. This was mostly likely due to small differences in the sensor setups of Experiment 1, which yielded training data for inducing the posture-based detector, and Experiment 3, which

served as the run-time test case for the posture-based detectors. The Kinect-based detectors were confirmed to work during piloting before the study, as well as after the study, but they were exceedingly conservative in detecting frustration and thus difficult to trigger. Further, the SVM classifier for detecting posture-based frustration was difficult to interpret, which made it challenging to identify postures and torso movements that would reliably trigger the detector. These factors limited the team's ability to adjust thresholds and increase the sensitivity of the posture-based frustration detectors prior to the study, particularly during the limited time that the team had access to the study room prior to the study.

As such, the Kinect-based condition essentially functioned as a control condition with no messages delivered to participants, rendering this part of the study a modified replication of Experiment 2. This unplanned replication was not an exact replication, as Experiment 2 had participants play five scenarios of TC3Sim, while Experiment 3 had participants only play three scenarios.

Results

The assumptions for normally distributed data for pre- and post-tests were met, as were assumptions for one-way ANOVA and rANOVA analyses. The BROMP measures of frustration were used as a covariate in the one-way ANOVA and rANOVA analyses. The determination to use this covariate was based on the results yielded in the second experiment that gave evidence there was an interaction effect of frustration on learning gains.

The first analysis run was to determine if there was a difference in learning gains between conditions. There were positive learning gains across all three conditions (see Fig. 8), but there was no significant difference in learning gains by condition, $F(2, 86) = .534$, $p = .588$, $\eta p^2 = .012$.

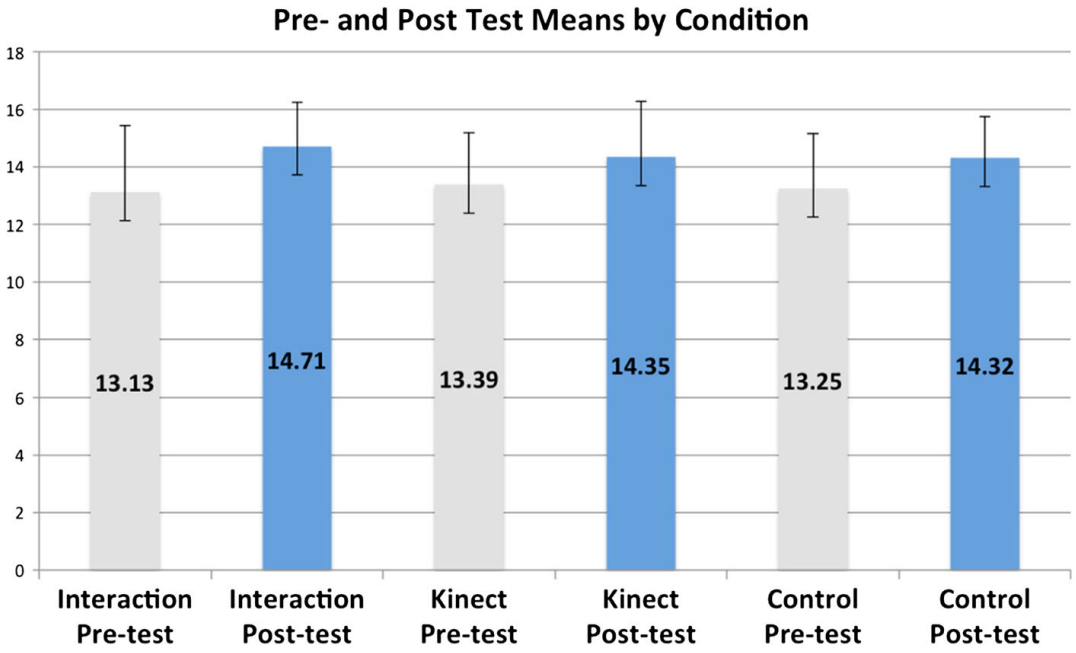

Fig. 8 Pre- and post-test means by condition

There was also no statistically significant difference in the frequency of field observations of frustration between all conditions, $F(2, 86) = 1.073$, $p = .346$, $\eta p^2 = .024$ (see Fig. 9). The highest proportion of frustration for any student in the control condition was 13.04%; the highest proportion of frustration for any student in the Kinect condition was 12.00%; the highest proportion of frustration for any student in the interaction detector condition was 10.00%. 38.2% of students in the control condition were never frustrated, 36.4% of students in the Kinect condition were never frustrated, and 38.2% of students in the interaction detector condition were never frustrated.

When adding the frequency of frustration field observations as a covariate in the rANOVA analysis, there was no statistically significant difference in learning gains between conditions, $F(1, 85) = .673$, $p = .513$, $\eta p^2 = .016$. Also, there were no significant interactions between BROMP frustration measures and condition, with learning gains $F(2, 83) = .231$, $p = .734$, $\eta p^2 = .007$.

Discussion of Affect Detector-Driven Frustration Feedback Study Results

This third experiment was designed to investigate whether there is a difference in efficacy between interventions administered according to a Kinect-based detector, an interaction-based detector and a fixed schedule designed to match the interaction-based detector in aggregate. However, the Kinect-based detector failed to trigger interventions, making the comparison actually between interventions administered according to an interaction-based detector, a fixed schedule designed to match the interaction-based detector in aggregate, and a no-intervention control. Our findings indicated that there was not a statistically significant difference on learning gains between these three conditions. Nor was there any statistically significant difference in frustration measured through BROMP, or an interaction effect between observed (BROMP) frustration measures and conditions on learning gains. These results contrast with the findings from Experiment 2, which found a significant effect of detector-triggered self-efficacy feedback messages on student learning gains.

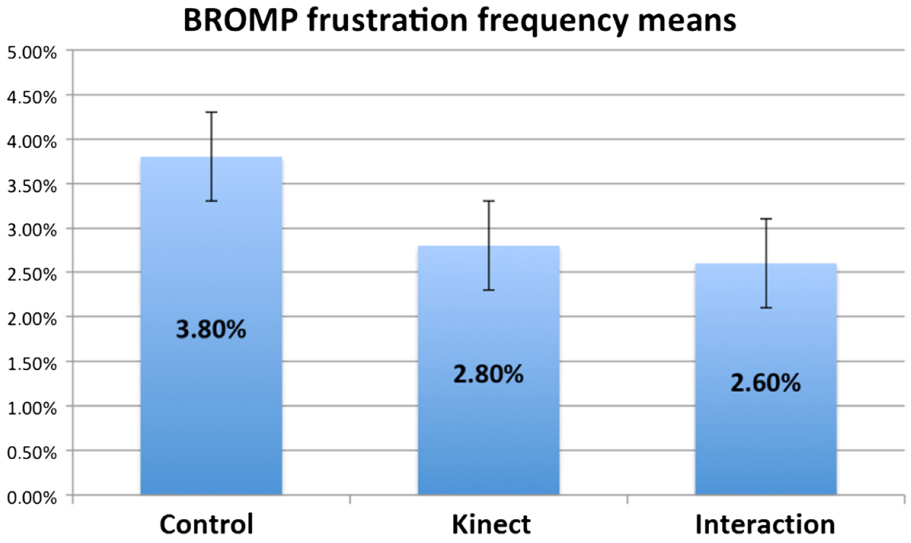

Fig. 9 BROMP observed frustration by condition: percentage means

While every student received a self-efficacy feedback message in the second and third scenarios in both the interaction-based condition (condition 1) and the fixed-schedule delivery condition (condition 3), these two conditions were not substantively different as to the timing of the delivery of the messages. The reason for this is that the control condition was a derivative of the interaction-based, sensor-free condition times calculated from the second experiment (September 2015) when the interaction-based detectors were employed. The mean times of delivered messages in the self-efficacy message condition, delivered upon the detection of frustration by the interaction-based frustration detectors, were calculated and these mean times were used to configure a fixed schedule of message delivery in this third experiment's control condition (condition 3). Students became frustrated at similar times in these scenarios designed to be frustrating, potentially reducing the usefulness of the automated detectors.

The results of this experiment also revealed that the Kinect-based frustration detectors did not get triggered in the second condition, and subsequently self-efficacy messages were not delivered to participants in this condition, rendering this essentially a second control condition. This condition was in turn associated with similar learning outcomes to the other condition. One possible explanation of the lack of difference between this control condition and the interaction-based condition is that this study represents a failed replication of the second experiment presented in this paper.

An alternate explanation is that the difference between the second and third experiment reflects the impact of a seemingly subtle difference in design between the two experiments. In specific, the second experiment was designed to be more frustrating than the first or third experiments, out of a concern that the protocol was overall insufficiently frustrating and a desire to magnify the impact of interventions while investigating different types of interventions. Specifically, in the second experiment, participants engaged in *five* scenarios—including two, no win *Kobayashi Maru* scenarios in between two scenarios that were actually winnable. The purpose behind this sequence was to undermine the expectations of success that the participants might have had and thereby increase the participants' frustration when they tried to win the two unwinnable scenarios.

This manipulation in the sequence of scenarios to elicit frustration suggests that frustration may not arise as a discrete event, but rather a form of affect that has a sustained, residual element. As such, the results imply that the lack of statistically significant findings between conditions in the third experiment may have been due to the choice of TC3Sim scenarios; perhaps the selected scenarios were insufficiently frustrating, thus making ameliorative interventions unnecessary. Yet another possible hypothesis is that the difference in results between the second and third experiments is simply the result of a dosage effect—five scenarios gave greater opportunity for the interventions to make an impact than three scenarios did.

Overall, several of these hypotheses suggest that the difference between Experiment 2 and Experiment 3 may have been due to the difference in scenarios the students experienced. As such, it seems reasonable to suggest that it would be valuable in the future to conduct a follow-up experiment comparing the self-efficacy interventions in the abbreviated scenario sequence (as in Experiment 3) to those same interventions in the extended scenario sequence (as in Experiment 2).

Conclusions and Future Work

This three year, multi-study project sought to develop interaction-based sensor-free affect detectors as well as sensor-based detectors, and then studying a range of interventions administered according to these detectors. After conducting a baseline study in September 2013 (experiment one), the results indicated that there was a negative correlation of frustration to learning gains while participants engaged in a modified TC3Sim, combat medic care training course. As such, the data collected from this baseline study was then used to develop the aforementioned affect detectors.

The second experiment successfully integrated the interaction-based detectors into TC3Sim by utilizing affect recognition functionalities provided by GIFT. This experiment specifically sought to identify whether motivational intervention feedback messages would yield greater learning gains than control conditions that did not have motivational feedback messages. The results of this second study demonstrated that there was a statistically significant difference between motivational and non-motivational conditions. Specifically, the self-efficacy feedback condition was found to be more effective than the other conditions.

Lastly, the third experiment sought to compare feedback messages triggered by the interaction-based, sensor-free detectors and the Kinect-based detectors. However, while the Kinect-based detectors did not detect participant frustration, the study yielded further data that contributes to the discussion of the complexity of identifying and addressing frustration in ITSs. Specifically, there was no difference in learning between interventions triggered by the interaction-based detector, interventions triggered on a fixed schedule, and no interventions at all, a seeming failure to replicate the second experiment. However, as discussed above, other differences in the design of the experiments may also explain this failed replication. Indeed, prior research has shown that the affective state of frustration is quite complex, where brief periods of frustration are not problematic while extended frustration is associated with worse learning outcomes (D'Mello and Graesser 2011; Liu et al. 2013; Robison et al. 2009). The second experiment (September 2015) and third experiment (March 2016) from this project support the notion that when addressing frustration, the duration and context of a frustrating event, as well as the nature of the interventions, cannot be casually addressed but rather require careful deliberation in the efforts to promote positive learning gains within ITSs.

There is much future work regarding the creation and validation of affective models, creation and validation of interventions, and better understanding how to use this type of intervention in practice. One lesson learned—not new to this effort but worth repeating—is that design efforts do not occur in a vacuum. Processing data from sensors and from within-environment interactions can serve as evidence for models. This model evidence can be refined through predictive metrics into student profile information. Student profile information can be used to trigger interventions, which can then be tested against other interventions to see which of them leads to better educational results, creating a single data point in a single domain for which approaches to addressing student affect are effective. The domain, the population, the design of the activity, the measures of learning, and the details of the detector feature engineering each influence the success of a specific intervention in a specific situation and context. There is certainly a large span of alternative interventions that could be considered for

this research; an alternate option, for example, would be to propose to the learner alternate steps that he or she could take to be successful in the simulation. While this would not actually help the learner in the *Kobayashi Maru* scenarios, it might create the perception that the player's success was under their control and reduce frustration (cf. Butterfield 1964).

The detectors presented here were moderately successful, but there was clearly room for improvement. Affect detector development is advancing as a field, but perhaps a bit more slowly and unevenly than one might hope. There are several possibly promising future direction. One potential area for future research, based on Brawner (2013), is using real-time model creation and query rather than creating a single group model for use on a number of individuals at a later time. Future affect detectors may now be compared to the models and results found here. Fundamentally, the use of the GIFT framework allows for the creation of a tutoring system in a variety of domains, integration of student modeling techniques, and the integration of RapidMiner-created models. The authors believe that the changes made to GIFT as part of this project—including support for integration between GIFT and RapidMiner—are valuable for future intelligent tutoring system creators to both develop and use.

The GIFT project is guided by the idea of creating generalized instructional strategies which map to specific instructional tactics. However, practical, specific instructional tactics must be used inside of the context of a specific domain. This work shows that, inside of the domain of TC3Sim, the strategy of swiftly responding to developed affective states may be useful under certain conditions to boost learning gains. The mechanism for how these messages had their impact is as yet uncertain; for example, they may have led to greater effort or cognitive engagement, in turn leading to the positive learning outcome seen. Beyond this, there are obviously other methods of instruction which may prove more valuable, better suited to other domains, or better suited to other categories of instruction. This task is a procedural task and it is somewhat unknown whether similar techniques would prove beneficial in a situational judgement of psychomotor type of task, or how modeling approaches can be changed in order to make the experience more beneficial. Similarly, a range of individual differences may play a major role in whether an intervention works for an individual learner.

The vast majority of the programming and program which are part of this effort have been made publicly available on www.gifttutoring.org. The first part of these changes is the availability of a Python plugin in order to perform data stream feature distillation in accordance with the features mentioned above. The second primary part of these changes is the creation of a plugin to RapidMiner, which is able to take the offline preprocessed data streams, make a model, and use that model at runtime in order to detect various emotional states. The number and types of features in addition to the number of types of states which can be acted on is limited only by the experimenter's ability to model.

Finally, GIFT provides placeholders for both recognized student states as well as "short term predicted" and "long term predicted" states. Eventually, it is anticipated that instructional actions can be taken *before* frustration/boredom is a realized state, rather than *after* it has been detected with confidence. The ability to take instructional actions based on *predicted future states* is in its relative infancy as a research study area, and is a valuable area of future study.

Acknowledgments This work is supported by U.S. Army Research Laboratory award, contract number W911NF-13-2-0008. We thank our research colleagues, COL James Ness and Dr. Michael Matthews in the Behavioral Science and Leadership Department at the United States Military Academy for their assistance in facilitating this research. Additionally, we would like to acknowledge the feedback and guidance of Drs. John Black and Dolores Perin of Teachers College, Columbia University, in regards to the design of Experiment 2. Any opinions, findings, and conclusions expressed in this paper are those of the authors and do not necessarily reflect the views of the U. S. Army.

Appendix: Feedback Messages For Experiment #2

Condition 1: Control-Value Theory

1. "Studies have shown that between 17%-19% of deaths in Vietnam could have been prevented if tourniquets had been used," (DePillis 2013).
2. "A 2008 study from a hospital in Baghdad found an 87% survival rate with use of tourniquets," (DePillis 2013).
3. "There is no room for hesitation or consultation in facial injuries, and quick action (3-10 minutes) is critical to the survival and recovery of injured soldiers," (Shuker 2011).
4. "The number one cause of preventable deaths in active shooter events is blood loss, and the best way to stop blood loss is to properly apply a tourniquet," (Jacobs et al. 2013).
5. "The first U.S. casualty to die in the war from enemy fire was a Special Forces Soldier, SFC Nathan Chapman, who died during medical air-evacuation on 4 January 2002 from isolated limb exsanguination without tourniquet use," (Kragh et al. 2013).

Condition 2: Social Identity Theory

1. As General Maxwell Thurman said, "Make good things happen for our Army."
2. Remember, solder, what General Patton said: "An Army is a team. It lives, sleeps, eats, and fights as a team."
3. "Every single man in this Army plays a vital role," said General Patton. "Don't ever let up. Every man has a job to do and he must do it."
4. General MacArthur once said: "Duty, Honor, Country, are three hallowed words that dictate what you ought to be, what you can be, what you will be."
5. General Patton said that the soldier is both a citizen and the Army, and the highest obligation and privilege of citizenship is the bearing arms for one's country.

Condition 3: Self-Efficacy Theory

1. In this important combat situation, your best outcomes will be achieved if you persist.
2. You can succeed in this because you've been trained to succeed under all conditions.
3. Tell yourself that you will succeed because failure is not an option in this high stakes combat zone.

4. Difficult doesn't mean impossible. It means work harder till your combat mission is achieved.
5. In all combat situations, success comes from overcoming the things you thought you couldn't.

Control Condition 1: Non motivational feedback messages

1. "Battlefield care emerged in Europe when Post-Revolutionary France established a system of prehospital care that included a corps of litter-bearers to remove wounded individuals from the battlefield," (Chapman et al. 2012).
2. "The modern combat medic has its roots in the American Civil War, when enlisted soldiers served as hospital stewards." (DeLorenzo 2001).
3. "As of 10 September 2001, the unreliable, World War II–era U.S. Army tourniquet was the only widely fielded tourniquet in the U.S. military," (Kragh et al. 2013).
4. "In 2003, in the farmlands around Fort Bragg, Amanda Westmoreland became a tourniquet maker by melting and bending plastic tourniquet components in her living room, packaging and distributing thousands of assembled tourniquets early in the war against Iraq," (Kragh et al. 2013).
5. "The use of a tourniquet went from a means of last resort to a means of first aid and became the prehospital medical breakthrough of the wars in Afghanistan and Iraq," (Kragh et al. 2013).

Control Condition 2: NO MESSAGES

References

AlZoubi, O., Calvo, R.A., & Stevens, R.H. (2009). Classification of EEG for emotion recognition: An adaptive approach. *Proceedings of the 22nd Australian Joint Conference on Artificial Intelligence*, pp. 52–61.

Amsel, A. (1992). Frustration theory: Many years later. *Psychological Bulletin, 112*(3), 396.

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

Arroyo, I., Woolf, B., Royer, J., Tai, M., Muldner, K., Burleson, W., & Cooper, D. (2010). *Gender matters: The impact of animated agents on students' affect, behavior and learning*, *Technical report UM-CS-2010-020*. UMASS Amherst: Computer Science Department.

Baker, R., Walonoski, J., Heffernan, N., Roll, I., Corbett, A., & Koedinger, K. (2008). Why students engage in" gaming the system" behavior in interactive learning environments. *Journal of Interactive Learning Research, 19*(2), 185.

Baker, R. S. (2007a). Modeling and understanding students' off-task behavior in intelligent tutoring systems. In Proceedings of the SIGCHI conference on human factors in computing systems (pp. 1059–1068). ACM.

Baker, R. S., Corbett, A. T., Koedinger, K. R., & Wagner, A. Z. (2004). Off-task behavior in the cognitive tutor classroom: when students game the system. In Proceedings of the SIGCHI conference on Human factors in computing systems (pp. 383-390). ACM.

Baker, R. S., D'Mello, S. K., Rodrigo, M. M. T., & Graesser, A. C. (2010). Better to be frustrated than bored: The incidence, persistence, and impact of learners' cognitive–affective states during interactions with

three different computer-based learning environments. *International Journal of Human-Computer Studies, 68*(4), 223-241.

Baker, R. S., & Ocumpaugh, J. (2015). *Interaction-based affect detection in educational software* (pp. 233–245). Oxford: The Oxford Handbook of Affective Computing.

Baker, R.S.J.d. (2007b). Is gaming the system state-or-trait? Educational data mining through the multi-contextual application of a validated behavioral model. *Complete on-line proceedings of the workshop on data Mining for User Modeling at the 11th international conference on user modeling* 2007, 76–80.

Baker, R. S. J. d., Gowda, S. M., & Corbett, A. T. (2011). Automatically detecting a Student's preparation for future learning: Help use is key. *Proceedings of the 4th International Conference on Educational Data Mining* (pp. 179–188).

Baker, R.S.J.d., Gowda, S. M., Wixon, M., Kalka, J., Wagner, A. Z., Salvi, A., Aleven, V., Kusbit, G., Ocumpaugh, J., & Rossi, L. (2012). Sensor-free automated detection of affect in a Cognitive Tutor for Algebra. In *Proceedings of the 5th International Conference on Educational Data Mining* (pp. 126–133).

Bandura, A. (1986). The explanatory and predictive scope of self-efficacy theory. *Journal of Social and Clinical Psychology, 4*(3), 359–373.

Bosch, N., D'Mello, S., Baker, R., Ocumpaugh, J., Shute, V., Ventura, M., et al. (2015). Automatic detection of learning-centered affective states in the wild. In Proceedings of the 20th international conference on intelligent user interfaces (pp. 379–388). ACM.

Brawner, K. (2013). Modeling learner mood in Realtime through biosensors for intelligent tutoring improvements. Electronic theses and dissertations. Paper 2608.

Butler, D. L., & Winne, P. H. (1995). Feedback and self-regulated learning: A theoretical synthesis. *Review of educational research, 65*(3), 245–281.

Butterfield, E. C. (1964). Locus of control, test anxiety, reactions to frustration, and achievement attitudes. *Journal of Personality, 32*(3), 355–370.

Calvo, R. A., & D'Mello, S. (2010). Affect detection: An interdisciplinary review of models, methods, and their applications. *IEEE Transactions on Affective Computing, 1*(1), 18–37.

Cetintas, S., Si, L., Xin, Y. P., Hord, C., Zhang, D. (2009). Learning to identify students' off-task behavior in intelligent tutoring systems. *Proceedings of the 14th International Conference on Artificial Intelligence in Education.*

Chapman, P. L., Cabrera, L. D., Varela-Mayer, C., Baker, M. M., Elnitsky, C., Figley, C., et al. (2012). Training, deployment preparation, and combat experiences of deployed health care personnel: Key findings from deployed US Army combat medics assigned to line units. *Military Medicine, 177*(3), 270–277.

Cocea, M., Hershkovitz, A., & Baker, R. S. J. d. (2009) The impact of off-task and gaming behaviors on learning: immediate or aggregate? *Proceedings of the 14th international Conference on artificial intelligence in education*, 507–514.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20*(1), 37–46.

Cooper, D. G., Arroyo, I., Woolf, B. P., Muldner, K., Burleson, W., & Christopherson, R. (2009). Sensors model student self concept in the classroom. In *International Conference on User Modeling, Adaptation, and Personalization* (pp. 30–41). Springer, Berlin.

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

D'Mello, S. K., & Kory, J. (2015). A review and meta-analysis of multimodal affect detection systems. *ACM Computing Surveys (CSUR), 47*(3), 43.

D'Mello, S. K., Craig, S. D., Witherspoon, A. W., McDaniel, B. T., & Graesser, A. C. (2008). Automatic detection of Learner's affect from conversational cues. *User Modeling and User-Adapted Interaction, 18*(1–2), 45–80.

D'Mello, S. K., Strain, A. C., Olney, A., & Graesser, A. (2013). Affect, meta-affect, and affect regulation during complex learning. In *International handbook of metacognition and learning technologies* (pp. 669–681). New York: Springer.

de Wit, S., & Dickinson, A. (2015). Ideomotor Mechanisms of Goal-Directed Behavior. In Braver, T. (2015). (Ed.), *Motivation and Cognitive Control* (pp. 143-163). London: Routledge.

DeLorenzo, R. A. (2001). Medic for the millennium: The US Army 91W health care specialist. *Military Medicine, 166*(8), 685.

DePillis, L. (2013). The return of the tourniquet: What we learned from war led to lives saved in Boston. *New Republic, April, 17*, 2013 Retrieved on February 8, 2015.

D'Mello, S., & Graesser, A. (2010). Mining bodily patterns of affective experience during learning. In *Proceedings of the 3rd International Conference on Educational Data Mining* (pp. 31–40).

D'Mello, S., & Graesser, A. (2011). The half-life of cognitive-affective states during complex learning. *Cognition & Emotion, 25*(7), 1299–1308.

Duckworth, A. L., & Quinn, P. D. (2009). Development and validation of the short grit scale (GRIT–S). *Journal of Personality Assessment, 91*(2), 166–174.

Frenzel, A. C., Pekrun, R., & Goetz, T. (2007). Girls and mathematics—A "hopeless" issue? A control-value approach to gender differences in emotions towards mathematics. *European Journal of Psychology of Education, 22*(4), 497–514.

Garcia-Ceja, E., Osmani, V., & Mayora, O. (2016). Automatic stress detection in working environments for smartphones' accelerometer data: A first step. *IEEE Journal of Biomedical and Health Informatics, 20*(4), 1053–1060.

Gjoreski, M. (2016). *Continuous stress monitoring using a wrist device and a smartphone*. Slovenia: Doctoral Dissertation, Jozef Stefan International Postgraduate School.

Goldberg, B., & Cannon-Bowers, J. (2015). Feedback source modality effects on training outcomes in a serious game: Pedagogical agents make a difference. *Computers in Human Behavior, 52*, 1–11.

Grafsgaard, J., Boyer, K., Wiebe, E., & Lester, J. (2012). Analyzing posture and affect in task-oriented tutoring. In *Proceedings of the 25th Florida Artificial Intelligence Research Society Conference*, 438–443.

Grafsgaard, J. F., Wiggins, J. B., Vail, A. K., Boyer, K. E., Wiebe, E. N., Lester, J. C. (2014). The additive value of multimodal features for predicting engagement, frustration, and learning during tutoring. In: *Proceedings of the 16th ACM International Conference on Multimodal Interaction* (pp. 42–49).

Hanley, J., & McNeil, B. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology, 143*, 29–36.

Hershkovitz, A., Wixon, M., Baker, R. S. J. d., Gobert, J., & Sao Pedro, M. (2011) Carelessness and Goal Orientation in a Science Microworld. Poster paper. *Proceedings of the 15th International Conference on Artificial Intelligence in Education* (pp. 462–465).

Hershkovitz, A., Baker, R. S. J., Gobert, J., Wixon, M., & Sao Pedro, M. (2013). Discovery with models a case study on carelessness in computer-based science inquiry. *American Behavioral Scientist, 57*(10), 1480–1499.

Hollands, F., & Bakir, I. (2015). *Efficiency of automated detectors of learner engagement and affect compared with traditional observation methods*. NY: Center for Benefit-Cost Studies of Education, Teachers College, Columbia University.

Jacobs, L. M., McSwain Jr, N. E., Rotondo, M. F., Wade, D., Fabbri, W., Eastman, A. L., et al. (2013). Improving survival from active shooter events: the Hartford Consensus. *Journal of Trauma and Acute Care Surgery, 74*(6), 1399-1400.

Jaques, N., Conati, C., Harley, J. M., & Azevedo, R. (2014). Predicting affect from gaze data during interaction with an intelligent tutoring system. In S. Trausan-Matu, K. Boyer, M. Crosby, & K. Panourgia (Eds.), *Proceedings of the 12th international conference on intelligent tutoring systems* (pp. 29–38). Heidelberg: Springer.

Kai, S., Paquette, L., Baker, R., Bosch, N., D'Mello, S., Ocumpaugh, J., et al. (2015). Comparison of face-based and interaction-based affect detectors in physics playground. In C. Romero, M. Pechenizkiy, J. Boticario, & O. Santos (Eds.), *Proceedings of the 8th International Conference on Educational Data Mining (EDM 2015)* (pp. 77–84). Pittsburgh: International Educational Data Mining Society.

Kapoor, A., Burleson, W., & Picard, R. W. (2007). Automatic prediction of frustration. *International Journal of Human-Computer Studies, 65*(8), 724–736.

Keselman, H. J. (1998) Testing treatment effects in repeated measures designs: An update for psychophysiological researchers. *Psychophysiology, 35*(4), 470-478

Klein, J., Moon, Y., & Picard, R. W. (2002). This computer responds to user frustration: Theory, design, and results. *Interacting with Computers, 14*(2), 119–140.

Kragh Jr., J. F., Walters, T. J., Westmoreland, T., Miller, R. M., Mabry, R. L., Kotwal, R. S., et al. (2013). *Tragedy into drama: An American history of tourniquet use in the current war*. TX: Army Inst of Surgical Research, Fort Sam Houston.

Liu Z, Pataranutaporn V, Ocumpaugh J, Baker RSJd (2013). Sequences of frustration and confusion, and learning. In: Proceedings of the 6th international conference on educational data mining, pp 114–120.

Mall, H., Goldberg, B. (2014) SIMILE: An authoring and reasoning system for GIFT. In *Proceedings of the 2nd Annual GIFT Users Symposium*. Orlando: US Army research laboratory.

Marien, H., Aarts, H., & Custer, R. (2015). How goals control behavior: The role of action- outcome and reward information. In Braver, T. (2015). (Ed.), Motivation and cognitive control (pp. 165–185). London: Routledge.

McQuiggan, S. W., & Lester, J. C. (2007). Modeling and evaluating empathy in embodied companion agents. *International Journal of Human-Computer Studies, 65*(4), 348–360.

Metcalf, S., Kamarainen, A., Tutwiler, M. S., Grotzer, T., & Dede, C. (2011). Ecosystem science learning via multi-user virtual environments. *International Journal of Gaming and Computer-Mediated Simulations, 3*(1), 86–90.

Mierswa, I., Wurst, M., Klinkenberg, R., Scholz, M., and Euler, T. (2006). YALE: Rapid prototyping for complex data mining tasks. *Proc. of the 12th ACM SIGKDD Int'l Conf. on Knowledge Discovery & Data Mining* (pp. 935–940).

Narciss, S. (2008). Feedback strategies for interactive learning tasks. *Handbook of research on educational communications and technology, 3*, 125–144.

Nasoz, F., Alvarez, K., Lisetti, C. L., & Finkelstein, N. (2004). Emotion from physiological signals using wireless sensors for presence technologies. *Cognition, Technology & Work, 6*, 4–14.

Ocumpaugh, J., Baker, R. S. J.d., & Rodrigo, M. M. T. (2012). *Baker-Rodrigo observation method protocol (BROMP). Training manual version 1.0. Technical report*. EdLab: New York, Manila: Ateneo Laboratory for the Learning Sciences.

Ocumpaugh, J., Baker, R. S., & Rodrigo, M. M. T. (2015a). *Baker Rodrigo Ocumpaugh monitoring protocol (BROMP) 2.0 technical and training manual. Technical report*. New York, Manila: Teachers College, Columbia University, Ateneo Laboratory for the Learning Sciences.

Ocumpaugh, J., Baker, R.S., Rodrigo, M.M.T., Salvi, A. van Velsen, M., Aghababyan, A., et al. (2015b). HART: The human affect recording tool. *Proceedings of the 33rd Annual International Conference on the Design of Communication (p. 24). ACM.*

Paquette, L., Baker, R. S., Sao Pedro, M. A., Gobert, J. D., Rossi, L., Nakama, A., et al. (2014). Sensor-free affect detection for a simulation-based science inquiry learning environment. In *Intelligent tutoring systems* (pp. 1–10). Springer International Publishing.

Paquette, L., Rowe, J., Baker, R.S., Mott, B., Lester, J., DeFalco, J., Brawner, K., Sottilare, R., Georgoulas, V. (2015). Sensor-Free or Sensor-Full: A Comparison of Data Modalities in Multi-Channel Affect Detection. Proceedings of the 8th International Conference on Educational Data Mining, 93–100.

Pardos, Z. A., Baker, R. S., San Pedro, M., Gowda, S. M., & Gowda, S. M. (2014). Affective states and state tests: Investigating how affect and engagement during the school year predict end-of-year learning outcomes. *Journal of Learning Analytic, 1*(1), 107–128.

Parsons, D. L., & Mott, J. (2005). 12-10 - tactical combat casualty care (TCCC) handbook. In *Fort Sam Houston, TX: Center for Army Lessons Learned*. Army Medical Department (AMEDD): United States Army.

Pekrun, R., Elliot, A., & Maier, M. (2006). Achievement goals and discrete achievement emotions: A theoretical model and prospective test. *Journal of Educational Psychology, 98*(3), 583–597.

Picard, R. W., Papert, S., Bender, W., Blumberg, B., Breazeal, C., Cavallo, D., et al. (2004). Affective learning—a manifesto. *BT technology journal, 22*(4), 253–269.

Robison, J., McQuiggan, S., & Lester, J. (2009). Evaluating the consequences of affective feedback in intelligent tutoring systems. In *Affective Computing and Intelligent Interaction and Workshops*, 2009. ACII 2009, *3rd International Conference on* (pp. 1–6). IEEE.

Rodrigo, M.M.T., Baker, R.S.J.d., McLaren, B., Jayme, A., and Dy, T. (2012). Development of a Workbench to Address the Educational Data Mining Bottleneck. *Proc. of the 5th Int'l Conf. on Educational Data Mining,* (pp. 152–155).

Roll, I., Aleven, V., McLaren, B. M., & Koedinger, K. R. (2011). Improving students' help-seeking skills using metacognitive feedback in an intelligent tutoring system. *Learning and Instruction, 21*(2), 267–280.

Rowe, J., Mott, B., McQuiggan, J., Robison, S., and Lester, J. (2009). Crystal Island: A narrative-centered learning environment for eighth grade microbiology. *Workshop on Educational Games at the 14th International Conference on Artificial Intelligence in Education,* (pp. 11–20).

Ryan, R. M. (1982). Control and information in the intrapersonal sphere: An extension of cognitive evaluation theory. *Journal of Personality and Social Psychology, 43*, 450–461.

Sabourin, J., Rowe, J., Mott, B., & Lester, J. (2011). When off-task in on-task: The affective role of off-task behavior in narrative-centered learning environments. In *Proceedings of the 15th International Conference on Artificial Intelligence in Education* (pp. 534–536).

San Pedro, M. O. C., Baker, R.S.J.d., Rodrigo, M. M. (2011a) The relationship between carelessness and affect in a cognitive tutor. In *Proceedings of the 4th bi-annual International Conference on affective computing and intelligent interaction* (pp. 306–315).

San Pedro, M. O. C., Baker, R., Rodrigo, M. M. (2011b) Detecting carelessness through contextual estimation of slip probabilities among students using an intelligent tutor for mathematics. In *Proceedings of 15th International Conference on Artificial Intelligence in Education* (pp. 304–311).

Sanghvi, J., Castellano, G., Leite, I., Pereira, A., McOwan, P. W., & Paiva, A. (2011). Automatic analysis of affective postures and body motion to detect engagement with a game companion. In *Proceedings of the 6th ACM/IEEE International Conference on Human-Robot Interaction* (pp. 305–311).

Schwarz, N. (2007). Attitude construction: Evaluation in context. *Social Cognition, 25*, 638–656.

Schwarz, N. (2009). Mental construal in social judgment. In F. Strack & J. Förster (Eds.), *Social cognition: The basis of human interaction* (pp. 121–138). New York: Psychology Press.

Shih, B., Koedinger, K., & Scheines, R. (2008). A response time model for bottom-out hints as worked examples. In *Proceedings of the 1st International Conference on Educational Data Mining* (pp. 117–126).

Shuker, S. (2011). The immediate lifesaving management of maxillofacial, life-threatening hemorrhages due to IED and/or shrapnel injuries: When hazard is in hesitation, not in the action. *Journal of Cranio-Maxillo-Facial Surgery, 40*, 534–540.

Slack, G. O. (2014). *The intersectionality between the short grit scale and four measures of academic self-efficacy of African American males enrolled through a remedial program at a historically black college and university*. Prairie View A&M University: Doctoral dissertation.

Smith, E., & Semin, G. (2004). Socially situated cognition: Cognition in its social context. *Advances in Experimental Social Psychology, 36*, 53–117.

Smith, E., & Semin, G. (2007). Situated social cognition. *Current Directions in Psychological Science, 16*, 132–135.

Sotomayor, T. M. (2010). Teaching tactical combat casualty care using the TC3Sim game based simulation: A study to measure training effectiveness. *Studies in Health Technology and Informatics, 154*, 176–179.

Sottilare, R. A., Brawner, K. W., Goldberg, B. S., & Holden, H. K. (2012). The Generalized Intelligent Framework for Tutoring (GIFT). Concept paper released as part of GIFT software documentation. Orlando, FL: US Army Research Laboratory–Human Research & Engineering Directorate (ARL-HRED). US Army Research Laboratory–Human Research & Engineering Directorate (ARL-HRED).

Tajfel, H., & Turner, J. C. (1979). An integrative theory of intergroup conflict. *The Social Psychology of Intergroup Relations, 33*(47), 74.

Verma, J. P. (2016). Repeated Measures Design for Empirical Researchers. Hoboken, NJ: John Wiley & Sons.

Wagster, J., Tan, J., Wu, Y., Biwas, G., & Schwartz, D. (2007). Do learning by teaching environments with metacognitive support help students develop better learning behaviors?. In Proceedings of the Cognitive Science Society (Vol. 29, No. 29).

Weinberg, S. L., & Abramowitz, S. K. (2002). Data Analysis for the Behavioral Sciences Using SPSS. New York: Cambridge University Press.

Witmer, B. G., & Singer, M. J. (1998). Measuring Presence in Virtual Environments: A Presence Questionnaire. *Presence: Teleoperators & Virtual Environments, 7*(3), 225-240. https://doi.org/10.1162/105474698565686.

Witmer, B. G., Jerome, C. J., & Singer, M. J. (2005). *The factor structure of the presence questionnaire. Presence, 14(3), 298 312*. Cambridge: MIT Press.

Wixon, M., Baker, R.S.J.d., Gobert, J., Ocumpaugh, J., & Bachmann, M. (2012). WTF? Detecting Students who are Conducting Inquiry Without Thinking Fastidiously. *Proceedings of the 20th International Conference on User Modeling, Adaptation and Personalization (UMAP 2012)*, (pp. 286–298).

Worsley, M., Scherer, S., Morency, L. P., & Blikstein, P. (2015). Exploring behavior representation for learning analytics. In *Proceedings of the 2015 ACM International Conference on Multimodal Interaction* (pp. 251–258).

Zeng, Z., Pantic, M., Roisman, G. I., & Huang, T. S. (2009). A survey of affect recognition methods: Audio, visual, and spontaneous expressions. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 31*(1), 39–58.