Modifying Field Observation Methods on the Fly: *Creative Metanarrative* and *Disgust* in an Environmental MUVE

Jaclyn Ocumpaugh[1], Ryan S. Baker[1], Amy M. Kamarainen[2], Shari J. Metcalf[3]

[1] Teachers College Columbia University, New York, New York
jo2424@tc.columbia.edu
[2] NY Hall of Science, New York, New York
[3] Harvard University, Massachusetts

Abstract. Automated detection of constructs associated with student engagement, disengagement, and meta-cognition plays an increasingly prominent part of personalized online education. Often these detectors are trained with ground truth labels obtained from field observations, a method that balances collection speed with label quality. Some behaviors and affective states (e.g., boredom) are regularly modeled across learning environments, but other constructs (e.g., gaming the system) manifest in fewer systems. New environments create the possibility of entirely unexpected constructs. In this paper, we describe how a field observation protocol (already proven effective for affect and behavior detection in several systems) was adapted to provide the flexibility needed to document previously unidentified or rare constructs. Specifically, we describe the in-field modification of the Baker Rodrigo Ocumpaugh Monitoring Protocol (BROMP) to accommodate categories not previously established (e.g., *creative metanarrative*) during observations of an educational multi-user virtual environment (MUVE). We also discuss the importance of developing methods that allow researchers to conduct such explorations while still capturing standard data constructs.

1 Introduction

As educational software has become more advanced, greater emphasis has been placed upon personalizing systems to react sensitively to student needs. Early work to model and adapt to student knowledge in tightly-scaffolded systems has given way to efforts to detect more ill-defined constructs (e.g. student engagement and metacognition) in more open-ended systems (e.g. virtual worlds). One approach to determining engagement with educational software is to construct automated detectors of affective states and behaviors, which can then be used both to research affect and learning [6, 8, 11] and to drive automated interventions [1,10].

Automated detectors have been produced from a variety of different data sources. Physical sensors (e.g., webcams, posture sensors and electroencephalograms) can be quite effective, but are often costly and fragile, making implementation difficult, particularly in poorer schools, leading some to develop sensor-free affect detection based on field observations [4, 15]. Recent research has expanded the scope of behavior detection to a wide range of systems, including games and simulations. As new environments are studied, we find that student behaviors differ across environments. Gaming the system is not seen in systems without feedback. WTF behaviors are more common in games than in tightly-constrained systems, and so on. As new systems are designed, fully anticipating relevant constructs may be impossible, particularly if classroom access or resources are limited. Given these concerns, researchers need coding methods that rigorously document known/expected constructs while being robust to unexpected findings is important.

In this paper, we discuss the adaption of the Baker Rodrigo Ocumpaugh Monitoring Protocol, (formerly the Baker-Rodrigo Observation Method Protocol), or BROMP, to address these concerns. BROMP is an established field observation method. It has been used to collect ground truth data for sensor-free models of affect and behavior and to study student engagement in non-technology-mediated learning environments. BROMP has already been successfully used to develop sensor-free affect detection in a variety of systems, including Cognitive Tutor [4] ASSISTments [15], and EcoMUVE [5] (the software described in this study). Here, we describe the expansion of BROMP coding schemes *in situ* to accommodate new affective and behavioral constructs that manifested during observations of EcoMUVE [13]. Currently, these constructs (*disgust* and *creative metanarrative*) are not typically coded for during field observations of educational software, but they will likely prove important as we increasingly rely upon virtual worlds for educational instruction.

2 Quantitative Field Observations (QFOs) using BROMP

The development of BROMP began in 2004 with field observations of students who were supposed to be learning from the Cognitive Tutor but were actually gaming the system [2, 3]. It was extended in 2007 when affective states were added as a second coding scheme [16], and further extended with the addition of teacher behaviors as well as student behaviors in some contexts [9]. In 2012, the method was formalized with the creation of a training manual [14]. New coders must achieve an adequate inter-rater reliability (Cohen’s Kappa of 0.6 for both affect and behavior, individually) with a trainer in order to become BROMP-certified. At present, 60 individuals have been certified for coding in the United States, the Philippines, and India.

BROMP works well for collecting ground truth observations of student affect and behavior both because of its simplicity and because the protocol is enforced by an app designed for Android, known as the Human Affect Recording Tool (HART) [4], which streamlines data collection process. At the beginning of each observation session, a coder inputs student login information into the HART application and selects a coding scheme. HART then presents each student’s login info back to the coder in the order in which they were entered. The coder then selects the behavioral and affect categories being presented by that student, ignoring the behaviors and affective states of other students except to the degree to which that information is contextually relevant to the student being coded.

3 BROMP Coding Schemes

During BROMP observations, behavior and affective states are coded separately but simultaneously. The coder has up to 20 seconds to categorize each student's behavior and affect, but records only the first thing he or she sees. For example, if a student is throwing a pencil at the teacher at the start of the observation, but then re-engages with the software while the coder is deciding what affective state is presenting, the behavior is recorded as off-task. In situations where a student has left the room, where the affect or behavior do not match any of the categories in the current coding scheme, or when the student can otherwise not be adequately observed, a '?' is recorded and that observation is eliminated from the data used to train automated detectors. This approach is valid when constructs that do not fit the coding scheme are rare, but researchers often need the flexibility to document new constructs.

The first BROMP publication to incorporate affect included seven different affective states and six behavioral categories [16]. These consisted of boredom, confusion, delight, surprise, frustration, flow, and neutral (drawn from [7]) as well as on task, on task conversation, off-task conversation, off-task solitary behavior, inactivity, and gaming the system (drawn from [2, 3]). However, at present, there are 24 coding schemes available, and it is possible to customize HART to a new schema.

The most commonly used BROMP coding schemes were developed for the Pittsburgh Science of Learning Center (PSLC). PSLC affective states include boredom, confusion, engaged concentration, frustration, and ?, while behavior categories include on task, on-task conversation, off-task, gaming the system, and ?. Because these constructs are seen as particularly relevant to educational settings, they are included in most BROMP coding schemes, but each time we work with a new learning environment, we reevaluate to ensure we are documenting all of the constructs relevant to that system and population.

4 Adapting BROMP Coding Schemes to EcoMUVE

When developing a coding scheme for EcoMUVE, expert field observers drew from prior coding schemes, from a qualitative pilot study, and from the EcoMUVE designers' expertise. We extended prior schemes with delight (which is seen substantially more in games than ITS) and sorrow, which is not typically included in educational research on affective states, being seen as rare [8]. We also extended the coding schema to enable us to document any categories that were unanticipated before entering the field, appending 3 different "user defined" categories (2 behavioral categories and 1 affective category) that the expert observer could specify in field.

Very early in the fieldwork, an affective state distinct from the anticipated categories emerged. As students began to explore this virtual world on day one, several reacted strongly EcoMUVE activities that they would have found "icky" in the real world, including tasks involving pond water or discoveries of dead fish. These reactions were coded as *disgust*, labeled as *User Defined 3* in HART. *Disgust* is rare in most learning, including EcoMUVE (0.04%) despite being one of Ekman's core emotions. Still, it was more prevalent than sorrow (0.03%), a category anticipated prior to fieldwork. Despite its negative valence, it indicates a lack of indifference. We do not yet know if it is positively or negatively associated with learning in EcoMUVE, but identifying this construct allows us to study how students respond to it. Anecdotally, students in this study maintained engagement once the *disgust* faded, but it could be an early indicator of later disengagement.

As fieldwork progressed, an unanticipated behavioral category was also identified. This behavior, which we term *creative metanarrative* (CM), was an unusual form of on-task conversation where students constructed their own storyline, often involving rogue police officers and illicit activities that did not reflect EcoMUVE design elements. CM differs from several other constructs that have been previously identified in the literature. While students often discuss content with each other during online learning (what [17] terms metanarrative), these students were transforming the plot of the game into a storyline that was more interesting to their peers. On it's face, this sounds similar to [12]'s transforming the game mechanic, which also includes a social component, or to previously identified WTF behaviors [18], but CM differs from these constructs behavior because it is not clear that it detracted from EcoMUVE's primary learning activities. In fact, the alternative storylines manufactured by these students may have made the software experience more exciting, forestalling the sort of unproductive within-game behaviors documented in [17, 18].

In contrast with the addition of *disgust* (which was coded within HART as soon as it was identified), the observer took more time to begin using the User Defined button in the behavioral coding scheme to code for *creative metanarrative*. This delay was driven by CM's relatively low frequency. Unlike *disgust*, CM did not manifest until the second day of field observation and only comprised 1.2% of the observations. (This is a low rate, but equal to the off-task behavior observed in this study.) Instead, the observer manually recorded this event on paper using the observation number and student number that HART provides as a reference at the top of each observation screen. After careful discussions with other BROMP-certified coders at the end of the second day of fieldwork, the field observer officially began automatically recording CM (using User Defined 1) in the field and the initial (manually recorded) instances were changed from the more generic on-task conversation to CM in the HART files.

5 Adapting BROMP Coding Schemes to EcoMUVE

Educational technology continues to evolve, and as it does researchers must have the tools that allow the agility to accurately and succinctly define relevant affective and behavioral constructs. As virtual worlds and other forms of educational software become more common educational tools, researchers are increasingly recognizing the importance of developing systems that are sensitive to indicators of student engagement. In particular, different systems promote different behavioral and affective responses. The quality and cost-effectiveness of field observation methods like BROMP make them an attractive option for collecting the ground truth labels needed for automated detectors of affect and behavior. In this paper, we discuss rapid, in-field extensions to BROMP (and HART, the software app used to enforce BROMP) that increase our ability to identify new constructs as we study student engagement in new software systems and populations.

Specifically, these extensions increase observers' agility to add unanticipated categories to the coding schemes in field, refining construct validity. While not correlated constructs, the two categories added in this study, *disgust* and CM, share qualities that are notable to educational researchers. Both manifest with rather prominent student displays within the classroom and may have broader impacts than their frequency would otherwise suggest. Both seem likely to reoccur in other virtual environments, suggesting that it may be increasingly important to take these constructs into account. Finally, both seem undesirable at first glance, but are actually indicators of engagement, suggesting that they may have interesting and complicated interactions with student outcomes. As researchers work to improve the sorts of engagement measures that facilitate the personalization of MUVEs, adding *disgust* and *creative metanarrative* to the suites of automated detectors already developed for systems like EcoMUVE [5] could substantially increase our understanding of learning and engagement, leading to greatly enhanced personalization options.

Acknowledgements

This research was supported by the Bill & Melinda Gates Foundation (#OPP1048577) and by the Institute of Education Sciences, U.S. Department of Education (#R305A080514).

References

1. Arroyo, I., Park Woolf, B., Cooper, D., Burleson, W., Muldner, K. (2011). The impact of animated pedagogical agents on girls' & boys' emotions, attitudes, behaviors & learning. *11th Int. Conf. Adv. Learning Tech. (ICALT)*, 506-510.
2. Baker, R.S., Corbett, A., Koedinger, K., Wagner, A. (2004). Off-Task Behavior in the Cognitive Tutor Classroom: When Students "Game The System." *Proc. ACM CHI: Computer-Human Interaction*, 383-390.
3. Baker, R.S., Corbett, A.T., Koedinger, K.R. (2004). Detecting Student Misuse of Intelligent Tutoring Systems. *Proc. 7th Int. Conf. Intell. Tutoring Sys*, 531-540.
4. Baker, R.S., Gowda, S.M., Wixon, M., Kalka, J., Wagner, A., Salvi, A., Aleven, V., Kusbit, G., Ocumpaugh, J., Rossi, L. (2012). Towards Sensor-free Affect Detection in Cog Tutor Algebra. *Proc. 5th Int. Conf. Ed. Data Mining*, 126-133.
5. Baker, R.S., Ocumpaugh, J., Gowda, S.M., Kamarainen, A., Metcalf, S. (in press) Extending Log-Based Affect Detection to a Multi-User Virtual Environment for Science. *Proc. 22nd Conf. User Modelling, Adaptation, & Personalization*.

6. 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.
7. D'Mello, S., Craig, S., Witherspoon, A., McDaniel, B., Graesser, A. (2005). Integrating affect sensors in an intelligent tutoring system. *Affective Interactions: The Comp. in the Affective Loop Wkshp, Int. Conf. Intell. User Interfaces,* 7-13.
8. D'Mello, S., Graesser, A. (2011). The half-life of cognitive-affective states during complex learning. *Cognition & Emotion* 25(7),1299-1308.
9. Godwin, K.E., Almeda, M.V., Petroccia, M., Baker, R.S., Fisher, A.V. (2013) Classroom activities and off-task behavior in elementary school children. *Proc. Annu. Meet. Cog. Sci. Soc.*, 2428-2433.
10. Lehman, B., D'Mello, S., Strain, A., Millis, C., Gross, M., Dobbins, A., Wallace, P., Millis, K., & Graesser, A. (2013). Inducing and tracking confusion with contradictions during complex learning, *Int. J. Artificial Intell. Ed.*, 22(2),85-105.
11. Liu, Z., Pataranutaporn, V., Ocumpaugh, J., Baker, R. (2013). Sequences of Frustration & Confusion, & Learning. *Proc. Int. Conf. Ed. Data Mining*, 114-120.
12. Magnussen, R., Misfeldt, M. (2004). Player Transformation of Educational Multiplayer Games. *Proc. Other Players.* Copenhagen, Denmark.
13. Metcalf, S., Kamarainen, A., Tutwiler, M.S., Grotzer, T., Dede C. (2011). Ecosystem science learning via multi-user virtual environments. *Int. J. Gaming & Computer-Mediated Simulations.* 3(1), 86.
14. Ocumpaugh, J., Baker, R.S.J.d., Rodrigo, M.A. (2012). The Baker-Rodrigo Observation Method Protocol (BROMP) 1.0 *Training Manual.*
15. Pardos, Z. A., Baker, R. S., San Pedro, M. O., Gowda, S. M., & Gowda, S. M. (2013). Affective states and state tests: Investigating how affect throughout the school year predicts end of year learning outcomes. *Proc. 3rd Int.Conf. Learning Analytics & Knowledge*, 117-124.
16. Rodrigo, M.M.T., Baker, R.S.J.d., Lagud, M.C.V., Lim, S.A.L., Macapanpan, A.F., Pascua, S.A.M.S., Santillano, J.Q., Sevilla, L.R.S., Sugay, J.O., Tep, S., Viehland, N.J.B. (2007). Affect and Usage Choices in Simulation Problem Solving Environments. *Proc. Artificial Intell. Ed.* 145-152.
17. Rowe, J. P., Shores, L. R., Mott, B. W., & Lester, J. C. (2010). A framework for narrative adaptation in interactive story-based learning environments. *Proc. of the Intell. Narrative Technologies III Workshop 14.*
18. Wixon, M., Baker, R.S.J.d., Gobert, J., Ocumpaugh, J., Bachmann, M. (2012). WTF? Detecting Students who are Conducting Inquiry Without Thinking Fastidiously. *Proc. Int. Conf. User Modeling, Adapt. & Personalization*, 286-298.