BROMP Quantitative Field Observations: A Review

Baker, R.S., Ocumpaugh, J.L., Andres, J.M.A.L

University of Pennsylvania

Abstract

In this chapter, we discuss over a decade of research to establish the Baker Rodrigo Ocumpaugh Monitoring Protocol (BROMP) as a method for conducting rapid, highly-quality, and time-synchronized quantitative field observations. We discuss work to establish standards and scalable training methods for the protocol in four countries, as well as work to develop and refine a handheld app that sped and facilitated uptake and use of the method. We then discuss the use of BROMP in research at scale: from its use in developing automated detectors of student engagement and affect, to use to study classroom pedagogy in settings from Pittsburgh to Chennai, to use in the iterative refinement of blended learning software.

Introduction

The systematic observation of classroom behavior is one of the most commonly-used methodologies in education research (Wilson & Reschly, 1996; Shapiro & Heick, 2004), and a variety of tools have been created to accommodate a range of research questions (see reviews in, for example, Adamson & Wachsmuth, 2014; Anderson, 1981; Fredericks et al., 2012; Hintz et al., 2002; Hops et al., 1995; Nock & Kurtz, 2005; Riley-Tillman et al., 2005; Skinner et al., 2000; Volpe et al., 2005). Like other systematic observation methods used in psychology (e.g., Furr & Funder, 2007), the direct observation of a classroom enables rigorous quantitative

analysis, as well-designed observation protocols (e.g., Hintz, 2005) reduce rater bias (Kent, O'Leary, Diament, & Dietz, 1974) and standardize data collection (Volpe and McConaughty, 2005) in studies of student and teacher behavior.

Classroom evaluation of teachers via observation protocols is used to assess instruction (Danielson, 2011; Junker et al., 2005; Walkington & Marder, 2013; PACT Consortium, 2012; Pianta, La Paro, & Hamre, 2008), with observation protocols addressing a range of research questions, including those which evaluate the content of instruction as well as those that evaluate pedagogical activities (Institute for Research on Policy Education and Practice, 2011) and their effect on student's learning and social outcomes (Grossman, Loeb, Cohen, & Wyckoff, 2013; Hill et al., 2012; Hamre, Goffin, Kraft-Sayre, 2009).

Many protocols attempt to capture large numbers of constructs across samples of students, focusing on overall aggregate frequencies or degrees of variables of interest. For example, the Reformed Teacher Observation Protocol (RTOP; Piburn et al. 2000, Sawada et al., 2002) uses Likert scales to evaluate the degree to which the teacher is encouraging certain kinds of student behaviors (e.g., communicative practices), and the Classroom Observation Scale (COS; Waxman & Padron, 2004) records student behaviors (e.g., on task, waiting for teacher, distracted, disruptive, or other) at a classroom level, along with teacher activities. Similarly, the Classroom Observation Protocol for Undergraduate STEM (COPUS; Smith, Jones, Gilbert, & Wieman, 2013) records a set of 12 instructor behaviors and 13 student behaviors, looking at the frequency of each behavior across 2-minute intervals, and the Code for Instructional Structure and Student Academic Response (CISSAR; Stanley and Greenwood, 1981; Carta et al., 1988) collects classroom observational data related to the classroom ecology, teacher behavior, and student behavior, sorted and categorized according to 105 individual codes in 13 categories of

variables (Lee, Wehmeyer, Palmer, Soukup, & Little, 2008; Soukup, Wehmeyer, Bashinski, & Bovaird, 2007). These tools measure a variety of variables at different grain-sizes and in different ways, but each is designed to collect data that can be used in both formative and summative analyses on the effect of different teacher practices on student engagement measures.

Observational protocols that focus on instructional practice generally do not log the frequency of student behavior for individual students, often for very practical reasons. It would not be feasible, for example, for a single observer to accurately report Likert scales on each student in a large classroom (e.g., the RTOP system) on an array of different measures, and introducing extra observers could quickly become a distraction that would invalidate any observations obtained. Frequency scores of very specific student behaviors (e.g., students raising their hands, a category included in COPUS), might be more manageable for a single observer, but could become unwieldy when trying to score each student for a large variety of codes, especially since there is documented bias toward identifying more extreme behaviors in aggregate classroom behavior rating scales (e.g., Christ et al., 2011).

Observations that focus on the behavioral patterns of individual students tend to be structurally and thematically different, both because of the challenges of observing multiple students and because many of them were developed for the screening and diagnosis of potential emotional and behavioral problems (Abikoff, Gittelman-Klein, & Klein, 1977; Gadow, Sprafkin, & Nolan, 1996; Walker & Severson, 1990). These protocols tend to focus on very small numbers of students at a time. For instance, the Classroom Observation Code (COC; Abikoff et al., 1977; Abikoff et al., 1980) and the Student Observation System (SOS; Reynolds & Kamphus, 2004) assess the classroom behaviors of students with ADHD in order to evaluate therapeutic interventions and the typicality of behavior of the diagnosed student. These studies often

evaluate only a small number of students at a time (e.g., two students in COC and one individual student in SOS.) The Individualized Classroom Assessment Scoring System (inCLASS; Downer et al., 2010) uses two observers to evaluate an individual pre-school student's self-regulatory skills.

Other observation systems for student behavior use a scan method for sampling, where the observer watches either a student or a group of students (sometimes the whole class) and talleys the presence of behaviors of interest as they occur. Richards et al. (2010) used a scan method to study the effect of interventions designed to to increase on-task behaviors. Harrison et al., (2014), however, argue that while observers using scanning methods are quite able to accurately rate broad categories on the more extreme ends of engagement (e.g., disruptive or compliant behaviors), they are less likely to achieve interrater reliability on less extreme behaviors. This finding reflects a general consensus among methodologists that scan-based observational methods are efficient but only appropriate for "overt, readily observable behaviors" (Ostrav and Heart, 2014).

While interrater reliability has been established for many observation systems, it has been insufficiently studied for many other methods (i.e., the discussion in Volpe et al., 2005). Some observation systems also require a considerable amount of training. While only a few hours is required for COS training (Waxman & Padron, 2004), 10-15 hours is required for the BOSS protocol (Shapiro, 2004; 2011; Volpe, DiPerna, Hintze, & Shapiro, 2005) and over 50 hours (across a three-week period) is reported to be required for COC training, which also has a high failure rate for coder certification (Abikoff et al., 1977; Abikoff et al., 1980). The time and resources consumed by longer training periods potentially limit the degree of applicability and scale which can be achieved for a method

The robust literature on classroom observations has provided us with considerable information that can be used to guide teacher training (Waxman & Padron, 1994) and diagnose student problems (Volpe et al., 2005), but for some time, there have been concerns that many of the observation systems have focused on categories that are convenient to code rather than those that are more connected to theoretical concerns (Ornstein, 1991; Lewis et al., 2014). While concrete categories like on/off-task behaviors can be quite important to learning (e.g., Karweit & Slavin, 1982), the need to explore student emotion and emotional supports, for example, has led to observation systems like the Classroom Assessment Scoring System (CLASS; Pianta, La Paro, & Hamre, 2008), which have expanded coding beyond more basic classifications of on/off task or compliance/disruption.

One area of particular recent interest has been affect related to student learning outcomes (Baker et al., 2010). Prior to the work described in this chapter and related work on video coding (e.g., Graesser et al., 2006), most of the quantitative research in affective learning employed some form of retrospective survey scales, either self-report (e.g., McCroskey's (1994) Affective Learning Scale, or ALS) or from an observer (e.g., Pianta et al.'s CLASS). Work around a decade ago began to instead follow Schutz and Pekrun's (2007) call for more observational measures of affective states to be developed, as these may capture more fluid and dynamic characteristics of affect than can be adequately described in post-hoc retrospective analyses. Systematic, direct observation of students' affective states may also provide opportunities to focus the grain size of the research to capture things that survey items may miss. For example, observational methods may help to contextualize the emotional experience (Fredericks et al., 2004), rather than ascribing it as a property of the individual, as some survey question formats may lend themselves to. (This criticism of surveys is given, for instance, in Wigfield and

Cambria, 2010). What's more, observations of emotion may serve to contextualize more traditional observational measures of behavior, as they may reveal the extent to which a student's active or passive compliance with an activity reveals their engagement with the material.

Early approaches to studying affect through video coding and observation often relied upon Ekman's Facial Action Coding System (FACS; Ekman & Friesen, 1978). Researchers using FACS code a set of six basic emotions, argued to be the only emotions seen universally across human cultures (Ekman & Friesen, 1971). However, more recent research suggests that many of these basic emotions are uncommon during learning, and that other, less universal, "academic" emotions are considerably more common in this context (D'Mello, Lehman, & Person, 2010). FACS has some key limitations; it requires hundreds of hours to learn (Ekman et al., 2002) and, as Furr & Funder (2007) note, even documenting the presence of a smile (a composite of several muscle units coded for in FACS) requires further interpretation, since a person may smile as the result of a variety of different emotional experiences. As such, interpreting emotions within this paradigm requires the use of a constellation of features (Planalp et al., 1998), a complex and time-consuming process that yields only a small number of emotions, many of which do not seem to be highly relevant to educational contexts.

The development of classroom observation systems that include codes for theoretically-grounded affective categories could help to expand the research community's understanding of the role of emotions in learning. However, these observation systems must be carefully constructed; in addition to considering the theoretical literature on emotions in learning and engagement (e.g., Csikszentmihali, 1990; D'Mello, Graesser, & Picard, 2007), they must also be sensitive to social and cultural differences between groups of students (e.g., Fischer et al., 2004; Krumhuber, Kappas, & Manstead, 2013). And, like more traditional behavioral observation

systems, observers of affective states must be aware of potential self-presentation effects and other observer effects (Wragg, 2002), particularly in cases where a student becomes sensitized to the observer's presence.

This chapter describes the development and use of a system that attempts to address these challenges, the Baker Rodrigo Ocumpaugh Monitoring Protocol (BROMP). It includes a description of BROMP's general practices and its adaptation for use in several different countries through the use of culturally sensitive training practices.

BROMP Overview and Theoretical Foundations

The Baker Rodrigo Ocumpaugh Monitoring Protocol (BROMP) was developed to address some of the limitations in prior classroom observation protocols, while also facilitating second-by-second synchronization between classroom observations and other data sources. BROMP is a method for conducting Quantitative Field Observations (QFOs) of student behavior and affect. BROMP is mostly used in classroom settings, but has been used in other settings as well (e.g. Baker et al., 2014). It was initially designed for observing students using educational software, however, BROMP has been adapted and used to study student engagement in other educational environments, such as research on elementary school students in classrooms with non-technological pedagogy (e.g. Hymavathy et al., 2014; Godwin et al., 2016).

Within BROMP, students are observed using a momentary time sampling method (Meany-Daboul, Roscoe, Bourret, & Ahearn, 2007), where students are repeatedly coded individually, in in a predetermined order. This sampling method is designed to achieve a representative sample of behavior and affect among the students while reducing the tendency to focus on more extreme events. Both affect and behavior are recorded in each observation. If a

student transitions between multiple affective states or behaviors within the 20 second interval, only the first affective state or behavior is coded. If two states seem to occur simultaneously, the observer is trained to code the most prominent behavior or affective state. If the behavior or affect presented by the student is outside of the coding scheme or if an observer is not sure what behavior or affect is being presented, a "?" is used to label the event. In the past, the "?" has also been used for cases where coding is impossible (e.g., the student is out of the room), but the protocol now differentiates students who are difficult to code (e.g., their affective state is ambiguous) from those who physically cannot be coded (e.g., they have been sent to the nurse's office). If a coder can make an assessment in less than 20 seconds, they do so and move on to the next student.

Observers record student affect and behavior using side glances and strategic positioning within the classroom. The frequency of observations recorded for each student is dependent on the number of students in the class. Coders consider multiple cues and environmental influences to which the students are exposed to (i.e., work context, actions, utterances, facial expressions, body language, and interactions with teachers or fellow students) in determining student affect and behavior (cf. Planalp et al., 1998).

BROMP assumes affect and behavior are at least partially orthogonal, and can be coded separately. Behavior and affect are coded simultaneously but separately using a custom-built Computer Assisted Direct Observation (CADO, Wessel, 2015) application for Android, the Human Affect Recording Tool (HART) (Ocumpaugh et al., 2015). HART presents coders with menus for affective or behavioral categories and automatically lists the potential categories that are currently being observed for. If required by the given research focus, HART also supports the

addition of a third coding scheme for recording classroom activities, teacher actions, and interventions in the classroom.

HART helps to enforce the correct use of the BROMP protocol in an automated fashion. It provides a far more precise time stamp than could possibly be manually recorded, synchronized to internet time, while enabling the observer to maintain their focus on the classroom context rather than the more mundane challenges of recording observations on paper. It also helps to reduce observer effects, as it is typically used on smaller handheld devices (e.g., smart phones, rather than tablets). This minimizes the number of observations being displayed on the screen at any given time (only the info for the student currently being observed), facilitating an observer's ability to collect information discreetly. HART also improves data security, both by limiting the information displayed and by ensuring that only someone with password-protected access is ultimately able to access the information.

HART's time stamps make it possible not only to sequence observations for analysis, but also make it possible to synchronize between HART data and other data streams. For example, it is possible to correlate a specific observation of a student with exactly what that student is doing within a learning application, either at the level of log files or actual screen recordings.

Over twenty coding schemes have been developed for a variety of learning environments in order to capture categories unique to certain environments and contexts. However, most incorporate a common set of academically-relevant emotions (e.g., Baker, D'Mello, Rodrigo, & Graesser, 2010), including boredom, confusion, engaged concentration, frustration, and delight (most commonly studied in game contexts) along with the codes for on-task or off-task student behavior. We will discuss some of the development of these coding schemes for different learning systems and countries below.

BROMP training and certification typically takes a few hours over the course of 1-2 days. BROMP coders are trained and certified through a process consisting three phases: 1) pre-field training, 2) field training, and 3) inter-rater reliability (IRR) testing. Pre-field training is often completed in one hour. In this phase, novices become familiar with the concepts that the field observations will be capturing. This is accomplished by reading the training manual, interacting with the android application, HART, and discussing the coding scheme and ambiguous examples with the trainer. Once this phase is completed, the trainee shadows an experienced coder in the field, discussing the coding process used for the observed students and practicing the use of the HART application. The field-training phase is most commonly accomplished within two or three observation sessions, lasting around 50 minutes per each. The last phase of BROMP certification is the IRR testing. In this field-testing phase, the novice's ability to conduct BROMP is field-tested using a criterion-referenced agreement test to determine their understanding of the process as well as the consistency of coding with those of the expert trainer (Ocumpaugh, Baker, & Rodrigo, 2015). During IRR testing, the novice and the trainer simultaneously code each student at the same time for at least 60 observations. If agreement is insufficient at the first IRR check, further field training occurs before testing for IRR again.

In order to be certified, trainees must first satisfy the recommended inter-rater agreement value. The BROMP measures IRR using Cohen's (1960) Kappa. Kappa scales between -1 to +1 and is used to reveal how much better the level of agreement is than what might be expected due to the base rate of each affective category. BROMP certification requires trainees to achieve a Kappa of 0.6 or higher for each coding scheme (both affect and behavior).

Another key aspect of BROMP training is discussion of how to minimize observer effects (e.g., McCall, 1984), including strategies for communicating with both teachers and students, as

well as techniques for the observer's positioning and movement, even in unusual classroom conditions. Observers coordinate with the teachers before the class observation begins to minimize any additional interaction throughout the sessions and to let the teachers know how to explain the observer's role to the students, if needed. During the observation session, observers are expected to avoid prolonged interaction with students without seeming threatening or interesting. Observers are recommended to maintain either a "clueless but friendly" or a "bored and disinterested" disposition (in contrast to some earlier protocols, where coders were instructed to be hostile to students to reduce student interest in interacting with them -- e.g. Reyes & Fennema, 1981). While recording, observers use side glances and peripheral vision. However, if a student becomes aware that they are being coded, the observation for the given time period is abandoned.

As McCall (1984) notes, the likelihood of observers creating effects on the data observed depends on the reactivity of the subjects, the social norms involved in the context being studied, and the degree to which the behaviors being studied are socially desirable. In our experience, US students are used to having outside observers come into classrooms to evaluate pedagogical practices, and since it is rare that these observations have any direct, salient consequences for students, the presence of BROMP coders is rarely questioned. Nevertheless, BROMP coders are also trained to minimize their presence in the classroom to remain as inconspicuous as possible, as affective states, in particular, are well-known to have social functions (e.g., Fischer & Manstead, 2008) that may be influenced by self-presentation, which has made BROMP observations more challenging in some contexts than in others. (See discussion in the section on BROMP Mexico, below.)

BROMP observers are discouraged from taking on other tasks during an observation session. The observer should not, for example, provide technical or pedagogical support, as this may sensitize the student to the observer. Similarly, if complicated coding schemes are required for other classroom activities (e.g., teacher behaviors), an additional observer is generally recommended. It is possible to keep track of simple changes in classroom activity, and the ability to code a tertiary coding scheme has been added to HART for such purposes. This functionality has been used in a few studies, including Ocumpaugh's (2016) study that denotes when the teacher transitions students from group work to solitary work, and in work by Godwin and colleagues (2016) and DiStefano (2018) to study how disengagement varies depending on which type of classroom activity is occurring. However, any tertiary scheme that divides the observer's attention away from the individual student being observed is likely to interfere with BROMP coding process, which can already be difficult in a crowded classroom.

BROMP USA

The first version of BROMP was developed as part of Ryan Baker's doctoral dissertation research. BROMP was first used, in its earliest form, in February 2003 and published in Baker, Corbett, Koedinger, & Wagner (2004). It used a momentary time sampling method to compare the effect of two different types of disengaged behavior during the use of educational software. Specifically, that research compared the learning associated with off-task behavior to the learning associated with gaming the system, a behavior where students progress through educational material by exploiting properties of the software (in this case systematic guessing or repeated quick help requests from the learning software) instead of by actively engaging in

learning. Researchers found that gaming the system was quite infrequent (3%) compared to other behaviors (78% for on-task solitary behavior, 4% on-task conversation, 1% for (active) off-task solitary behavior, 11% for off-task conversation, and 3% for inactivity), but it had a disproportionate effect on learning, correlating to poorer learning twice as strongly as off-task behavior. Papers soon followed where BROMP observations were used to inform the development of automated models of gaming the system (Baker, Corbett, & Koedinger, 2004) and off-task behavior (Baker, 2007). BROMP was also used to assess the effectiveness of automated interventions attempting to reduce the frequency of gaming the system (Baker et al., 2006). In this early work, involving only disengaged behavior, an inter-rater reliability of 0.84 (Cohen's Kappa) was achieved.

In these early studies, BROMP observations included only observations of behavior and were recorded using pen and paper. The first observations of BROMP involving affect were conducted in the Philippines in 2006, and are discussed in that section. In 2009, researchers moved from recording BROMP observations with pen and paper to using HART, discussed above. HART went through several design revisions and upgrades, primarily developed in the context of BROMP USA.

In 2012, the first BROMP manual (Ocumpaugh et al., 2012) was developed in the United States, but in conjunction with researchers in the Philippines, as part of efforts to streamline the training process. This effort formalized the training process, increased methodological consistency between the researchers in the two countries where it was being used, and cut the US training time down from 10-15 hours across 4-5 days to 5-7 hours across 1-3 days (sometimes shorter). It was later expanded upon in a more comprehensive manual (Ocumpaugh et al., 2015),

which broadened the effort to more thoroughly document the theoretical foundations underlying the observation system.

BROMP has been extensively used in the United States, where there are currently more than 50 certified coders. As these numbers have grown, BROMP-based research has expanded beyond its original use (Baker et al., 2004) in the development of automated models of student behaviors in educational software. In addition to continuing to be used to model student behaviors (e.g. Pardos et al., 2014; Jiang et al., 2018), it has now been used to develop automated models of student emotions in many types of learning systems (Baker et al., 2012; Miller et al., 2014; Baker et al., 2014; Bosch et al., 2016; Jiang et al., 2018; Gobert et al., 2015). These automated "detectors" are developed by synchronizing field observations to the exact behavior occurring in the learning system at the same time, using the precise time stamps provided by the HART handheld app.

In turn, the detectors built for these systems have been used to do study how different patterns of affective states relate to learning outcomes, including some non-traditional long-term measures. For example, they have been used to model state standardized exam scores (e.g., Pardos et al., 2014), college attendance (San Pedro et al., 2013), and enrollment in STEM careers or graduate degree programs post-college (Patikorn et al., 2018). These detectors have also been used to conduct more basic research on the relationship between patterns of different affective states and behaviors in learning contexts (e.g., Liu et al., 2013; Karumbaiah et al., 2018; Ocumpaugh et al., 2017; Baker et al., 2007; Rodrigo et al., 2012; Baker et al., 2011; Bothelo et al., 2018).

BROMP has also been used to conduct research in non-technological contexts. For example, Godwin and colleagues used BROMP to study how classroom design impacts student engagement in kindergarten and elementary school classrooms (Godwin et al., 2016), and DiStefano (2018) used BROMP to study engagement in graduate-level courses. Others have used BROMP in unpublished studies to provide formative engagement measures to teachers, within a regional charter school network, and to study the affect of learners participating in informal science education programs.

BROMP Philippines

BROMP in the Philippines began in 2006 when researchers were seeking an alternative to sensor-based detection of student affect and behavior. Sensors commonly used in wealthier nations, such as galvanic skin response sensors and posture sensors, are neither readily available in the Philippines nor economically feasible options for local research. After meetings between Didith Rodrigo and Ryan Baker (the first developer of BROMP in the US), Rodrigo's team in the Philippines started conducting data collection using a protocol inspired by Baker's work in detecting gaming behavior among students (discussed above), but extended the protocol to include a second coding scheme (simultaneously implemented) for affective states, focusing on flow (engaged concentration), confusion, frustration, boredom, neutral, delight, and surprise.

The initial version of the modified protocol was tested in a private school in Quezon City, Metro Manila and was initially implemented by either two or three coders who observed individual students together. As part of scaling BROMP in the Philippines, a more formal certification process for new coders was adopted, requiring an inter-rater reliability (Cohen's

Kappa) of 0.6 or higher -- a number lower than was common in the highly straightforward coding schemes popular in many laboratory psychology studies, but much higher than the numbers seen for other affect labeling protocols such as FACS (Sayette et al., 2001).

In this early work, observations were conducted in pairs or groups of three either once per per 180 seconds (Andallaza, Rodrigo, Lagud, Jimenez, & Sugay, 2012; Rodrigo & Baker, 2011; Rodrigo et al., 2012; San Pedro, Rodrigo, & Baker, 2011), , and the coders observed each student within a fixed period of 20 seconds. More recent studies (Andres et al., 2015; Banawan, Rodrigo, & Andres, 2015) have moved to the American practice of having a single observer conduct the protocol, and have adjusted the observation length to fall between 5 and 8 seconds, when faster labeling is feasible.

Within the Philippines, BROMP has been used for a range of analyses. Some of the first studies using BROMP in the Philippines included the first-ever study showing the long-term correlations between moment-to-moment affect and learning outcomes (Rodrigo et al., 2009), the second-ever study of affective dynamics (Baker, Rodrigo, & Xolocotzin, 2007), and the first-ever study comparing affect in intelligent tutors and games (Rodrigo et al., 2008). Later BROMP research in the Philippines examined the relationship of confusion with student driven in-game events (Andres et al., 2014), the relationship of student affect with the frequency of careless errors (San Pedro, Baker, & Rodrigo, 2014), the differences between the facial expressions, behavior, and confusion trajectories of Filipino and American learners (Tabanao & Rodrigo, 2016), the correlations of boredom and confusion with student action sequences (Andres et al., 2015), the relationship between frustration and student achievement (Banawan, Rodrigo, & Andres, 2015), the affective experiences around figuring out a challenging problem (Andres et al., 2014), and the development of a predictive model on wheel-spinning (Palaoag et al., 2016).

All in all, BROMP has been used in the Philippines by 17 BROMP-certified coders, throughout the country (from Manila to Davao), and in well over a hundred published scientific articles.

BROMP India

In 2014, India became the third country to have BROMP-certified coders. The process of seeding BROMP for India was conducted by two local researchers, in consultation with Baker, who traveled to India to participate in the process. To understand affect and behavior in India to adapt BROMP for use in India, the two Indian researchers conducted qualitative observations in multiple classrooms in government-run schools. Baker initially intended to participate in this process, but his presence was highly disruptive to the these classrooms. The Indian researchers met with Baker outside the school context to develop a first-draft coding scheme, and implement it within the HART phone app. They then returned to the school to practice coding together before checking for inter-rater reliability. The researchers discussed the results before returning to the classroom to try again. The two Indian coders reached acceptance inter-rater reliability on their second try, with a Kappa of 0.78 for behavior and 0.72 for affect. After this point, the two initial coders trained and certified a third coder in the following week, and then the three of them began training other coders over the following months.

One of the key goals in India was to train a large number of coders for broad use of BROMP. Therefore, efforts were made to further streamline the training process. A video was created by the American team explaining the overall process and theory behind BROMP coding. This video was shown to groups of coders, and was followed of an explanation of the process and question-and-answer session held locally. After this, the same certification process (of

simultaneous but separate coding of the same student and the same time by a previously-certified coder and a new coder, requiring Kappa over 0.6) was followed. A total of 72 people were certified in BROMP within India; 68 of them teachers or school personnel in a single large city, working with a local research center. These coders used a slightly-modified version of the HART app, which adjusted the interface to fit very small and inexpensive phones.

The coding scheme adopted was different in India than in the USA or Philippines, based on the different affect and behavior expressed by students in India and in the learning context. Behaviors such as gaming the system were not relevant in the Indian classrooms; instead three behaviors were coded: active participation (42.7% during last seeder inter-rater checking session), passive participation (41.1% during last seeder inter-rater checking session), and off-task (16.4% during last seeder inter-rater checking session). Boredom (34.3%), confusion (4.1%), and engaged concentration (34.3%) were seen in both India and earlier settings; other affective states were new to India, specifically disinterest (1.4%), enthusiasm (8.2%), and mild interest (17.8%). Engaged concentration was referred to as "focused" to be clearer to the coders, but was considered to be the same category by the researchers. A final category, contempt, was hypothesized as a counterpart to the American practice of coding for frustration, but was not seen during the seeder inter-rater checking session.

BROMP was used in India for multiple purposes. It was used to provide formative data to teachers and school leaders on individual students' engagement (Hymavathy, Krishnamani, & Sumathi, 2014). It was used to determine which students might be at risk of dropping out (Hymavathy, Krishnamani, & Sumathi, 2015), and it was used to assess whether a specific reform curricula was more effective at engaging students than traditional classroom practice.

However, the systematic and extensive use of BROMP in India has largely ceased after governmental changes and the dissolution of the scientific research center that used BROMP.

BROMP UK

BROMP was adapted for use in the UK in 2015 by researchers at the University of London, led by Manolis Mavrikis, Beate Grawemeyer, and Wayne Holmes. These researchers needed a systematic protocol to gather affective and behavioral observations of students for studies on adaptive feedback, and used BROMP is several studies (e.g. Grawemeyer, Mavrikis, Holmes, Hansen, Loibl, & Gutiérrez-Santos, 2015a; Grawemeyer, Mavrikis, Holmes, Gutierrez-Santos, 2015b; Grawemeyer et al., 2016; Grawemeyer, Mavrikis, Holmes, Gutiérrez-Santos, Wiedmann, & Rummel, 2017).

Research in the UK focused on younger students (usually 8-10 years old) who were using the iTalk2Learn math tutor. In early studies (Grawemeyer et al., 2015), Interrater Reliability (IRR) was quite good (Cohen's Kappa = 0.71), but was established using a substantially different method than implemented in other countries. In this classroom study, BROMP field-workers recorded several affective states (boredom, confusion, frustration, surprise, and flow/concentration) as well as behavioral states (off task, on task, on task conversation, on task reflection, and gaming the system). The BROMP observers' codes were then compared to codes generated by researchers who retrospectively coded both video and audio recordings of the classroom study after the fact (Grawemeyer et al., 2015).

Later, a subset of BROMP-generated field data was also compared to automated detectors of affect (Grawemeyer et al., 2017). The automated detectors were informed by

students' interactions with the iTalk2Learn software (including voice recordings of students' interactions with a help system). In this research, the reliability between BROMP observers and automated detectors of affect was more moderate (Kappa = .551), in part because the automated detectors never recognized an instance of delight (Kappa = .643 with delight excluded).

BROMP UAE

BROMP was adapted for use in the United Arab Emirates in 2018 by researchers affiliated with Alef Education, working in public schools, led by Aishah Al Yammahi, Rand Muhsen, and Guzelle Shahid. There was some consideration of adding an additional affective state of restlessness, but it was not found to be common within classrooms, within the initial investigations. The commonly occurring affective states in the UAE were the frequently-studied boredom, frustration, confusion, engaged concentration, and delight. A new behavioral category of distracted was also added to represent students who were multi-tasking while actively working within the learning system. Each of the three initial BROMP coders achieved inter-rater reliability with each other of above 0.6, for both affect and engaged/disengaged behavior. BROMP norming for Abu Dhabi was only completed immediately before publication of this article, but there are plans to use BROMP both for curricular refinement and comparison of different curricula and designs.

BROMP Mexico

As discussed in previous sections, BROMP has been successfully implemented outside of the United States in four other countries: the Philippines, India, England, and the United Arab

Emirates. A team of researchers led by Genaro Rebolledo-Mendez at the Universidad Veracruzana attempted to bring BROMP to Mexico as well. Rebolledo-Mendez planned the adaptation of BROMP to Mexico during an extended visit Baker and Ocumpaugh's laboratory in the United States. Having previously used other methods to study affect and behavior, specifically self reports and video observation, BROMP provided an alternative that only required minimal technical setup.

In 2015, a pilot of BROMP was conducted in Veracruz, Mexico, within a public secondary school. In this study, a set of 83 students were observed while interacting with a Cognitive Tutor for middle school mathematics that had previously been studied using BROMP in the United States (e.g., Baker et al., 2004). Across several 50 minute class sessions, multiple Mexican researchers (accompanied by researchers from the USA) attempted to create a coding scheme and then reach acceptable inter-rater reliability. However, acceptable inter-rater reliability was not reached despite multiple attempts conducted by three different Mexican researchers from the same region as the students being observed.

One possible explanation is that students may have been uncomfortable displaying visible affect in the presence of the external observers and their teachers due to cultural factors not seen in the other countries where BROMP has been established. It's possible, for example, that the high power distance (the degree of hierarchical inequality which is acceptable between a superior and subordinate--in this case, the teacher and students (Bochner & Hesketh, 1994) may have contributed to this phenomenon. Some evidence against this hypothesis, however, is the success of BROMP in the Philippines, which Hofstede's (1984) research suggests has slightly higher power distance than Mexico (Hofstede, 1984). For example, Hofstede (1983) reports a substantial difference between Mexico and the Philippines/USA for a separate, but related

cultural dimension known as uncertainty avoidance, which capture the degree to which individuals are comfortable with ambiguous or uncertain situations, such as those seen in the use of a new and very different practice during learning. Hofstede and McCrae (2004) suggest that uncertainty avoidance is related to the expression or inhibition of affect, and Mesquita and Walker's (2003) research on cross-cultural differences in the regulation of emotions suggests that children with high levels of uncertainty avoidance, as is common in Mexico, may seek to avoid demonstrating undesirable behaviors or affect. These values may lead to self-presentation effects (e.g., suppressing emotions with negative valance), which could make it more difficult for outside observers who are interested in affective states like confusion, boredom, and frustration. The researchers noted that students seemed concerned about being evaluated on criteria that was unknown to them and as a result remained constantly aware of the researchers noting their observations. Unfortunately, there is not yet evidence available to disambiguate between these possible explanations.

Whatever the reason, the first attempt at implementing BROMP in Mexico was unsuccessful. Future attempts to study BROMP in Mexico are currently in the planning stages, and researchers will study the potential hypotheses for BROMP's initial failure in future attempts to refine BROMP for use in Mexico. A protocol like BROMP cannot succeed if students are uncomfortable with the observation process. It may be possible to determine whether the difficulty in observation is due to observer effects by coding the same categories from video data rather than field observation data. It also may be possible to reduce observer effects by more clearly explaining to students that observations are for research purposes only and will not impact their grades. Overall, the process of adapting BROMP for successful use in Mexico may not solely be beneficial in terms of making BROMP available for research; it may produce

broader insight into the characteristics of students in Mexico and the conditions of Mexican classrooms.

Summary of Cross-cultural Development Standards and Research

The development of BROMP coding schemes for multiple countries offers the opportunity for new kinds of research on student engagement, but only if the adaption of BROMP for new cultures adheres to appropriate safeguards. In addition to adhering to strict training and interrater reliability checks and taking precautions to ensure that observer effects are appropriately minimized, there are two key components to developing BROMP in new cultural environments that are absolutely essential.

First, while some consistency in the content of BROMP coding schemes is important for comparative studies, it would be highly inappropriate to fully impose an American coding scheme in other contexts. Affective categories have different meanings in different contexts, and these differences need to be taken into account. For example, researchers in India who collaborated with us on the development of coding schemes for use there suggested that "frustration" would be seen as a socially embarrassing category that would be subject to considerable presentation effects, and therefore it was not included in BROMP-India coding schemes. Therefore, any development of BROMP for a new cultural context must be done in consultation with researchers who are native to the culture being observed.

However, despite the need to revise coding schemes for different cultural contexts, one interesting finding is how much commonality there is across countries. Boredom and confusion show up in every country we have normed BROMP for. Engaged concentration also shows up in every country we have normed BROMP for, though an alternate term "focused" was preferred in

India. Delight is seen in every country studied when game-based or playful contexts are considered. Frustration is seen in 4 of the 5 countries where BROMP has been normed, with the exception in India discussed in the previous paragraph. Behaviors also have a great deal of commonality between countries, with general off-task and on-task categories being common across countries, though other constructs are more prominent in some contexts than others. For example, on-task conversation with other students is relatively rarer in the Philippines than in the other countries where BROMP has been studied, and multi-tasking "distracted" behavior appears to be more common in the UAE than in the other contexts studied, leading to its inclusion in UAE coding schemes but not in other countries. Still, the difference between learning contexts (e.g. games versus intelligent tutors versus teacher-led instruction) appears to be at least as important in terms of the behavior seen as the cultural context is.

Second, the judgments of non-native researchers should not serve as the base-line in interrater reliability checks, even if that researcher has considerable BROMP-coding experience in his or her own culture. While there is some evidence in the literature about the universality of six basic emotions (e.g., Ekman & Friesen, 1971), there are growing concerns about these claims (e.g., Nelson & Russell, 2013; Jack et al., 2012; Porter & Samovar 1996; Tsai & Levenson, 1997; Matsumoto, 1991), and the affective states typically studied with BROMP are more complex than the basic emotions claimed to be universal (Ortony, Clore, & Collins, 1990). Moreover, given the holistic approach that people use to perceive and identify emotions (e.g., Panalp et al. 1998), it is not surprising that research shows that these perceptions are not culturally universal (e.g., Gendron et al., 2014). There is also the possibility that self-presentation effects may be magnified for academic and achievement-related emotions. This is not to say that all BROMP coders must be from the same population as they are coding -- indeed, at the time of

this writing four people have become BROMP-certified for cultures they did not grow up in -- but that the baseline for certifying BROMP coders and establishing protocols should rest on native coders. Typically, cross-cultural coding is both less reliable and contains specific biases (Okur et al., 2018).

Meeting these conditions makes it possible to use BROMP in a valid and useful way in new contexts. Failing to meet these conditions creates significant risks of invalid inference, and researchers conducting affect coding in new contexts without meeting these standards are not using BROMP.

BROMP in the broader learning science context

Recent decades have seen increased interest in engagement and affect within among learning science researchers (Reschly & Christenson, 2012). One of the key recent trends in learning science has been the increasing awareness of the value of larger-scale data, both in terms of more intensive data, more data per learner (cf. Blikstein, 2011), and data on more learners (Baker & Siemens, 2014). Traditional methods for measuring engagement and affect, such as self-report, emote-aloud, video coding, and previous generations of classroom observation protocols, are limited in terms of the degree to which they can achieve both intensity and scale. BROMP makes it feasible to achieve higher intensity and scale compared to previous-generation methods for measuring engagement and affect. In addition, by facilitating the development of automated detectors of engagement and affect (e.g. Baker et al., 2012, 2014; DeFalco et al., 2018; Kai et al., 2015; Miller et al., 2014; Pardos et al., 2014; Bosch et al., 2016; Jiang et al., 2018; Gobert et al., 2015), BROMP makes it feasible to use this even more intensive and scalable method.

The value of BROMP to the broader endeavor of learning science can be seen in the broad range of research questions and problems that have become more amenable to analysis, through BROMP. This is particularly true for research on affect. The majority of studies on the moment-by-moment prevalence of different affective states during learning have been conducted using BROMP. The majority of studies on affective dynamics have been conducted using BROMP. Well more than half of the studies of the relationship between moment-by-moment affect and student outcomes have either involved BROMP or automated detectors developed using BROMP. By far, the largest-scale studies on student affect have involved BROMP or automated detectors developed using BROMP (e.g. Pardos et al., 2014; Godwin et al., 2016). A large proportion of modern research on affect in the learning sciences has been made possible through BROMP.

BROMP has not been quite as dominant a method for engagement and disengagement research, but has still played a substantial role, leading to papers that have been cited, in aggregate, well over a thousand times, on phenomenon ranging from studying the relationship between disengagement and learning (e.g. Baker, Corbett, Koedinger, & Wagner, 2004) to studies on the disengagement associated with different pedagogies (Mulqueeny et al., 2015) to studies on the relative proportion of different forms of disengagement in different countries (Rodrigo, Baker, & Rossi, 2013).

As engagement and affect continue to be key foci of learning science research, we anticipate that methods like BROMP will continue to be an important part of research in this area. The trend in learning science is towards research methods that are scalable but with established validity, a trend in line with BROMP's advantages.

Conclusions and ongoing development

BROMP was originally developed for the training of automated models of student engagement among students using educational software in the US. The protocol led to the first automated detector of gaming the system, a type of disengaged behavior that has since been shown to correlate with poorer student outcomes. Relatively quickly, in consultation with Didith Rodrigo, its use was expanded from an observation protocol used solely to code for behavior to one that included a separate coding scheme for affective states (and since then, for other classroom conditions as well). Since 2012, when it was used for the first sensor-free, automated models of student of students' affective states (Baker et al., 2012), BROMP has been used to develop automated detectors of engagement and affect for several learning systems (Baker et al., 2012, 2014; DeFalco et al., 2018; Kai et al., 2015; Miller et al., 2014; Pardos et al., 2014; Bosch et al., 2016; Jiang et al., 2018; Gobert et al., 2015).

BROMP's use has expanded considerably over the last decade. Currently, BROMP has been used to study more than 20 different software systems, and there are over 150 BROMP-certified coders in the US, the Philippines, India, the UK, and the UAE. Done properly, there is no reason to think BROMP could not be adapted to a range of other international contexts as well, although the example of Mexico shows that doing so may sometimes present unexpected challenges.

In addition to its use in other countries, BROMP has increasingly been used in non-technological contexts. The extension of its use, from training educational software to recognize how students behave when they are engaged (or not) to estimating the frequency of behaviors and affective states in a range of contexts, has also led to new research questions. However, there is not currently sufficient knowledge about how large samples need to be for momentary time

sampling methods such as BROMP. While extremely large studies are clearly safe (e.g. Godwin et al., 2016), we do not yet know how small our samples can safely be. Further statistical work to develop power analysis for this context (e.g., Paquette et al., 2015) is needed.

Additionally, BROMP may support research into understanding what thresholds are important for intervention during the development of affect. For example, a student may need an intervention for confusion eventually, but understanding what length of confusion or frustration is educationally desirable in a given context requires further research. Likewise, understanding the antecedents to boredom (Baker et al., 2011) is important to developing appropriate interventions (e.g., Dashmann et al., 2011; Pekrun et al., 2010).

However, there is now a growing body of research showing interest in both the kinds of automated models that BROMP is used to construct (cf., D'Mello, Dieterle, & Duckworth, 2017) and in the use of BROMP for other types of classroom observations. BROMP has also influenced the development of new approaches to coding student interaction and experiences, including the Human Expert Labeling Protocol (HELP) developed by Intel (Aslan et al., 2017), and the system used by researchers at Carnegie Mellon to develop the Spatial Classroom Log Explorer (SPACLE), a system that helps to integrate analysis of different types of learning data (Holstein et al., 2017). We look forward to seeing how this research unfolds in the coming years.

References

Abikoff, H., Gittelman, R., & Klein, D. F. (1980). Classroom observation code for hyperactive children: A replication of validity. *Journal of Consulting and Clinical Psychology, 48*(5), 555.

Abikoff, H., Gittelman-Klein, R., & Klein, D. F. (1977). Validation of a classroom observation code for hyperactive children. *Journal of Consulting and Clinical Psychology*, *45*(5), 772.

Achenbach, T. M. (1986). *The Direct Observation Form*. University of Vermont, Department of psychiatry.

Adamson, R. M., & Wachsmuth, S. T. (2014). A review of direct observation research within the past decade in the field of emotional and behavioral disorders. *Behavioral Disorders, 39*(4), 181-189.

Andallaza, T. C. S., Rodrigo, M. M. T., Lagud, M. C. V., Jimenez, R. J. M., & Sugay, J. O. (2012). Modeling the Affective States of Students Using an Intelligent Tutoring System for Algebra. In *Proceedings of The Third International Workshop on Empathic Computing (IWEC 2012)* (3),4.

Anderson, L. W. (1981). Instruction and time-on-task: A review. J. *Curriculum Studies, 13*(4), 289-303.

Andres, J. M. L., Rodrigo, M. M. T., Baker, R. S., Paquette, L., Shute, V. J., & Ventura, M. (2015, June). Analyzing Student Action Sequences and Affect While Playing Physics Playground. In *EDM (Workshops)*.

Aslan, S., Mete, S. E., Okur, E., Oktay, E., Alyuz, N., Genc, U. E., ... & Esme, A. A. (2017). Human expert labeling process (HELP): towards a reliable higher-order user state labeling process and tool to assess student engagement. *Educational Technology*, 53-59.

Baker, R.S.J.d. (2007) Modeling and Understanding Students' Off-Task Behavior in Intelligent Tutoring Systems. *Proceedings of ACM CHI 2007: Computer-Human Interaction*, 1059-1068.

Baker, R.S.J.d., Corbett, A.T., Koedinger, K.R., Evenson, S.E., Roll, I., Wagner, A.Z., Naim, M., Raspat, J., Baker, D.J., Beck, J. (2006) Adapting to When Students Game an Intelligent Tutoring System. *Proceedings of the 8th International Conference on Intelligent Tutoring Systems*, 392-401.

Baker, R.S.J.d., 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., 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., Corbett, A. T., & Koedinger, K. R. (2004). Detecting student misuse of intelligent tutoring systems. In *International Conference on Intelligent Tutoring Systems* (pp. 531-540). Springer, Berlin, Heidelberg.

Baker, R.S.J.d., Moore, G., Wagner, A., Kalka, J., Karabinos, M., Ashe, C., Yaron, D. (2011) The dynamics between student affect and behavior occurring outside of educational software. *Proceedings of the 4th bi-annual International Conference on Affective Computing and Intelligent Interaction*.

Baker, R.S., Ocumpaugh, J., Gowda, S.M., Kamarainen, A., Metcalf, S.J. (2014) Extending log-based affect detection to a multi-user virtual environment for science. *Proceedings of the 22nd Conference on User Modelling, Adaptation, and Personalization*, 290-300.

Baker, R.S.J.d., Rodrigo, M.M.T., Xolocotzin, U.E. (2007) The dynamics of affective Transitions in simulation problem-solving environments. *Proceedings of the Second International Conference on Affective Computing and Intelligent Interaction* .

Baker, R., Siemens, G. (2014) Educational data mining and learning analytics. In Sawyer, K. (Ed.) *Cambridge Handbook of the Learning Sciences: 2nd Edition*, pp. 253-274.

Banawan, M. P., Rodrigo, M. M. T., & Andres, J. M. L. (2015). Investigation of frustration among students using Physics Playground. *Proceedings of the 23rd International Conference for Computers in Education.*

Berry III, R. Q., Rimm-Kaufman, S. E., Ottmar, E. M., Walkowiak, T. A., & Merritt, E. G. (2013). The mathematics scan (M-Scan): A measure of standards-based mathematics teaching practices. *Unpublished measure). Charlottesville, VA: University of Virginia.*

Blikstein, P. (2011). Using learning analytics to assess students' behavior in open-ended programming tasks. In *Proceedings of the 1st international conference on learning analytics and knowledge* (pp. 110-116). ACM.

Bochner, S., & Hesketh, B. (1994). Power distance, individualism/collectivism, and job-related attitudes in a culturally diverse work group. *Journal of cross-cultural psychology, 25*(2), 233-257.

Bosch, N., D'Mello, S., Baker, R., Ocumpaugh, J., Shute, V., Ventura, M., ... & Zhao, W. (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.

Bosch, N., D'mello, S. K., Ocumpaugh, J., Baker, R. S., & Shute, V. (2016). Using video to automatically detect learner affect in computer-enabled classrooms. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, *6*(2), 17.

Carta, J. J., Greenwood, R., Schulte, D., Arreaga-Mayer, C., & Terry, B. (1988). *Code Instructional Structure and Student Academic Response: Mainstream Version (MS-CISSAR)*. Kansas City: Juniper Gardens Children's Project, Bureau of Child Research, University of Kansas.

Christ, T. J., Riley-Tillman, T. C., Chafouleas, S., & Jaffery, R. (2011). Direct Behavior Rating: An Evaluation of Alternate Definitions to Assess Classroom Behaviors. *School Psychology Review*, *40*(2).

Csikszentmihalyi, M. (1990). *Flow: The psychology of optimal performance*.

Daschmann, E. C., Goetz, T., & Stupnisky, R. H. (2011). Testing the predictors of boredom at school: Development and validation of the precursors to boredom scales. *British Journal of Educational Psychology*, *81*, 421–440.

Danielson, C. (2011). *Enhancing professional practice: A framework for teaching*. ASCD.

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

DiStefano, D. (2018). *How Pre-Service Teachers' Engagement and Affect Informs Instructional Format of an Introductory Methods Course.* Unpublished doctoral dissertation, Fordham University.

D'Mello, S. K., Lehman, B., & Person, N. (2010). Monitoring affect states during effortful problem solving activities. *International Journal of Artificial Intelligence in Education*, *20*(4), 361-389.

D'Mello, S., Dieterle, E., & Duckworth, A. (2017). Advanced, analytic, automated (AAA) measurement of engagement during learning. *Educational psychologist*, *52*(2), 104-123.

D'Mello, S., Picard, R. W., & Graesser, A. (2007). Toward an affect-sensitive AutoTutor. *IEEE Intelligent Systems, 22*(4).

Downer, J. T., Booren, L. M., Lima, O. K., Luckner, A. E., & Pianta, R. C. (2010). The Individualized Classroom Assessment Scoring System (inCLASS): Preliminary reliability and validity of a system for observing preschoolers' competence in classroom interactions. *Early Childhood Research Quarterly*, *25*(1), 1-16.

Ekman, P., & Friesen, W. V. (1971). Constants across cultures in the face and emotion. *Journal of Personality and Social Psychology*, *17*(2), 124.

Ekman, P., & Friesen, W. V. (1978). *Facial Action Coding System: Investigator's Guide*. Consulting Psychologists Press.

Ekman, P., Friesen, W. V., & Hager, J. C. (Eds.). (2002). *Facial Action Coding System* [E-book]. Salt Lake City, UT: Research Nexus.

Fischer, A. H., Rodriguez Mosquera, P. M., Van Vianen, A. E., & Manstead, A. S. (2004). Gender and culture differences in emotion. *Emotion, 4*(1), 87.

Fredricks, J.A., Blumenfeld, P.C., Paris, A.H. (2004). School engagement: Potential of the concept, state of the evidence. *Review of Educational Research,74*, 59 –109.

Fredricks, J. A., & McColskey, W. (2012). The measurement of student engagement: A comparative analysis of various methods and student self-report instruments. In *Handbook of Research on Student Engagement* (pp. 763-782). Springer.

Frick, T., & Semmel, M. I. (1978). Observer agreement and reliabilities of classroom observational measures. *Review of Educational Research*, *48*(1), 157-184.

Furr, R. M., & Funder, D. C. (2007). Behavioral observation. *Handbook of Research Methods in Personality Psychology*, 273-291.

Gadow, K. D., Sprafkin, J., & Nolan, E. E. (1996). *Attention Deficit Hyperactivity Disorder School Observation Code*.

Gendron, M., Roberson, D., van der Vyver, J. M., & Barrett, L. F. (2014). Perceptions of emotion from facial expressions are not culturally universal: evidence from a remote culture. *Emotion*, *14*(2), 251.

Gobert, J. D., Baker, R. S., & Wixon, M. B. (2015). Operationalizing and detecting disengagement within online science microworlds. *Educational Psychologist*, *50*(1), 43-57.

Godwin, K.E., Almeda, M.V., Seltman, H., Kai, S., Skerbetz, M.D., Baker, R.S., Fisher, A.V. (2016). Off-task behavior in elementary school children. *Learning and Instruction,* 44, 128-143.

Graesser, A. C., McDaniel, B., Chipman, P., Witherspoon, A., D'Mello, S., & Gholson, B. (2006). Detection of emotions during learning with AutoTutor. In *Proceedings of the 28th annual meetings of the cognitive science society* (pp. 285-290).

Grawemeyer, B., Mavrikis, M., Holmes, W., Hansen, A., Loibl, K., Gutiérrez-Santos, S. (2015). Affect matters: Exploring the impact of feedback during mathematical tasks in an exploratory environment. In *International Conference on Artificial Intelligence in Education*. Springer, Cham. 595-599.

Grawemeyer, B., Mavrikis, M., Holmes, W., Hansen, A., Loibl, K., & Gutiérrez-Santos, S. (2015). The impact of feedback on students' affective states. In *CEUR Workshop Proceedings (Vol. 1432)*. CEUR Workshop Proceedings.

Grawemeyer, B., Mavrikis, M., Holmes, W., & Gutierrez-Santos, S. (2015). Adapting feedback types according to students' affective states. In *International Conference on Artificial Intelligence in Education*. Springer, Cham. pp. 586-590)

Grawemeyer, B., Mavrikis, M., Holmes, W., Gutierrez-Santos, S., Wiedmann, M., & Rummel, N. (2016). Affecting off-task behaviour: how affect-aware feedback can improve student learning. In *Proceedings of the sixth international conference on learning analytics & knowledge* (pp. 104-113). ACM.

Grawemeyer, B., Mavrikis, M., Holmes, W., Gutiérrez-Santos, S., Wiedmann, M., & Rummel, N. (2017). Affective learning: improving engagement and enhancing learning with affect-aware feedback. *User Modeling and User-Adapted Interaction, 27*(1), 119-158.

Grossman, P., Loeb, S., Cohen, J., & Wyckoff, J. (2013). Measure for measure: The relationship between measures of instructional practice in middle school English language arts and teachers' value-added scores. *American Journal of Education, 119*(3), 445-470.

Hamre, B. K., Goffin, S. G., & Kraft-Sayre, M. (2009). *Classroom Assessment Scoring System Implementation Guide: Measuring and Improving Classroom Interactions in Early Classroom Settings*.

Harrison, S. E., Riley-Tillman, T. C., & Chafouleas, S. M. (2014). Direct behavior rating: Considerations for rater accuracy. *Canadian Journal of School Psychology*, *29*(1), 3-20.

Hill, H.C., Charalambous, C.Y., Blazar, D., McGinn, D., Kraft, M.A., Beisiegel, M., Humez, A., Litke, E. and Lynch, K. (2012). Validating arguments for observational instruments: Attending to multiple sources of variation. *Educational Assessment, 17*(2-3), 88-106.

Hintze, J. M. (2005). Psychometrics of direct observation. *School Psychology Review, 34*(4), 507.

Hofstede, G. (1984). *Culture's Consequences: International Differences in Work-Related Values* (Vol. 5). Sage.

Hofstede, G. (2003). *Culture's Consequences: Comparing Values, Behaviors, Institutions and Organizations Across Nations.* Sage publications.

Hops, H., Davis, B., & Longoria, N. (1995). Methodological issues in direct observation: Illustrations with the living in familial environments (LIFE) coding system. *Journal of Clinical Child Psychology*, 24, 193–203.

Holstein, K., McLaren, B. M., & Aleven, V. (2017). SPACLE: investigating learning across virtual and physical spaces using spatial replays. In *Proceedings of the Seventh International Learning Analytics & Knowledge Conference* (pp. 358-367). ACM.

Hymavathy, C., Krishnamani, V. R., & Sumathi, C. P. (2014). Analyzing learner engagement to enhance the teaching-learning experience. In MOOC, Innovation and Technology in Education (MITE), *2014 IEEE International Conference*. 67-70.

Hymavathy, C., Krishnamani, V. R., Sumathi, C. P. (2015). EDM: Finding answers to reduce dropout rates in schools. *International Journal of Computing Algorithm*, *4*(1), 26-29.

Institute for Research on Policy Education and Practice. (2011). *Protocol for Language Arts Teaching Observation (PLATO).* http://platorubric.stanford.edu/

Jack, R. E., Garrod, O. G., Yu, H., Caldara, R., & Schyns, P. G. (2012). Facial expressions of emotion are not culturally universal. *Proceedings of the National Academy of Sciences*, *109*(19), 7241-7244.

Jiang, Y., Bosch, N., Baker, R. S., Paquette, L., Ocumpaugh, J., Andres, J. M. A. L., ... & Biswas, G. (2018). Expert Feature-Engineering vs. Deep Neural Networks: Which Is

Better for Sensor-Free Affect Detection?. In *International Conference on Artificial Intelligence in Education* (pp. 198-211). Springer, Cham.

Junker, B. W., Weisberg, Y., Matsumura, L. C., Crosson, A., Wolf, M., Levison, A., & Resnick, L. (2005). *Overview of the instructional quality assessment*. Regents of the University of California.

Kai, S., Paquette, L., Baker, R.S., Bosch, N., D'Mello, S., Ocumpaugh, J., Shute, V., Ventura, M. (2015) A Comparison of Video-based and Interaction-based Affect Detectors in Physics Playground. *Proceedings of the 8th International Conference on Educational Data Mining* , 77-84.

Karweit, N., & Slavin, R. E. (1982). Time-on-task: Issues of timing, sampling, and definition. *Journal of Educational Psychology, 74*(6), 844.

Kent, R. N., O'leary, K. D., Diament, C., & Dietz, A. (1974). Expectation biases in observational evaluation of therapeutic change. *Journal of Consulting and Clinical Psychology, 42*(6), 774.

Krumhuber, E. G., Kappas, A., & Manstead, A. S. (2013). Effects of dynamic aspects of facial expressions: A review. *Emotion Review, 5*(1), 41-46.

Lee, S. H., Wehmeyer, M. L., Palmer, S. B., Soukup, J. H., & Little, T. D. (2008). Self-determination and access to the general education curriculum. *The Journal of Special Education, 42*(2), 91-107.

Lewis, T. J., Scott, T. M., Wehby, J. H., & Wills, H. P. (2014). Direct observation of teacher and student behavior in school settings: Trends, issues and future directions. *Behavioral Disorders, 39*(4), 190-200.

Lund, T. J., Pilarz, M., Velasco, J. B., Chakraverty, D., Rosploch, K., Undersander, M., & Stains, M. (2015). The best of both worlds: Building on the COPUS and RTOP observation protocols to easily and reliably measure various levels of reformed instructional practice. *CBE—Life Sciences Education*, *14*(2), ar18.

Matsumoto, D. (1991). Cultural influences on facial expressions of emotion. *Southern Journal of Communication*, *56*(2), 128-137.

McCall, G. J. (1984). Systematic field observation. *Annual review of sociology*, *10*(1), 263-282.

McCoy, S., Galletta, D. F., & King, W. R. (2005). Integrating national culture into IS research: The need for current individual level measures. *Communications of the Association for Information Systems, 15*(1), 12.

McCroskey, J. C. (1994). Assessment of affect toward communication and affect toward instruction in communication. In *1994 SCA Summer Conference Proceedings and Prepared Remarks: Assessing College Student Competence in Speech Communication.* Annandale, VA: Speech Communication Association.

Meany-Daboul, M. G., Roscoe, E. M., Bourret, J. C., & Ahearn, W. H. (2007). A comparison of momentary time sampling and partial-interval recording for evaluating functional relations. *Journal of Applied Behavior Analysis*, *40*(3), 501-514.

Mesquita, B., & Walker, R. (2003). Cultural differences in emotions: A context for interpreting emotional experiences. *Behaviour Research and Therapy*, *41*(7), 777-793.

Miller, W. L., Petsche, K., Baker, R. S., Labrum, M. J., & Wagner, A. Z. (2014). Boredom across activities, and across the year, within reasoning mind. In *Workshop on Data Mining for Educational Assessment and Feedback (ASSESS 2014).*

Mulqueeny, K., Kostyuk, V., Baker, R.S., Ocumpaugh, J. (2015) Incorporating Effective e-Learning Principles to Improve Student Engagement in Middle-School Mathematics. *International Journal of STEM Education, 2* (15).

Nelson, N. L., & Russell, J. A. (2013). Universality revisited. *Emotion Review*, *5*(1), 8-15.Nock, M. K., & Kurtz, S. M. (2005). Direct behavioral observation in school settings: Bringing science to practice. *Cognitive and Behavioral Practice, 12*(3), 359-370.

Ocumpaugh, J., d Baker, R. S., Gaudino, S., Labrum, M. J., & Dezendorf, T. (2013). Field observations of engagement in Reasoning Mind. *International Conference on Artificial Intelligence in Education*. Springer, Berlin, Heidelberg. 624-627.

Ocumpaugh, J. (2015). *Baker Rodrigo Ocumpaugh Monitoring Protocol (BROMP) 2.0 Technical and Training Manual.* New York, NY and Manila, Philippines: Teachers College, Columbia University and Ateneo Laboratory for the Learning Sciences.

Okur, E., Aslan, S., Alyuz, N., Esme, A.A., Baker, R.S. (2018). Role of socio-cultural differences in labeling students' affective states. *Proceedings of the 19th International Conference on Artificial Intelligence in Education,* 367-380.

Ornstein, A. C. (1990). A look at teacher effectiveness research—theory and practice. *NASSP Bulletin, 74*(528), 78-88.

Ortony, A., Clore, G. L., & Collins, A. (1990). *The Cognitive Structure of Emotions*. Cambridge university press.

Ostrov, J. M., & Hart, E. (2014). Observational methods. *The Oxford handbook of quantitative methods in psychology: foundations*, *1*, 286.

Paquette, L., Ocumpaugh, J., & Baker, R. S. (2015). Simulating Multi-Subject Momentary Time Sampling. In *EDM.* 586-587.

PACT Consortium (2012). *PACT (Performance Assessment for California Teachers)*. http://www.pacttpa.org/_main/hub.php?pageName=Home

Palaoag, T. D., Rodrigo, M. M. T., Andres, J. M. L., Andres, J. M. A. L., & Beck, J. E. (2016). Wheel-spinning in a game-based learning environment for physics. In *International Conference on Intelligent Tutoring Systems* (pp. 234-239). Springer, Cham.

Pardos, Z.A., Baker, R.S., San Pedro, M.O.C.Z., 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 Analytics, 1* (1), 107-128

Patikorn, T., Heffernan, N.T., Baker, R.S. (2018). ASSISTments Longitudinal Data Mining Competition 2017: A Preface. *Proceedings of the Workshop on Scientific Findings from the ASSISTments Longitudinal Data Competition, International Conference on Educational Data Mining*.

Pekrun, R., Goetz, T., Daniels, L., Stupnisky, R. H., & Perry, R. (2010). Boredom in achievement settings: Exploring control–value antecedents and performance outcomes of a neglected emotion. *Journal of Educational Psychology*, *102*, 531–549.

Pianta, R. C., La Paro, K. M., & Hamre, B. K. (2008). *Classroom Assessment Scoring System™: Manual K-3*. Paul H Brookes Publishing.Piburn, M., & Sawada, D. (2000). *Reformed Teaching Observation Protocol (RTOP) Reference Manual.* Technical Report.Picard, R. W., Papert, S., Bender, W., Blumberg, B., Breazeal, C., Cavallo, D., ... & Strohecker, C. (2004). Affective learning—a manifesto. *BT technology journal, 22*(4), 253-269.Planalp, S. (1998). Communicating emotion in everyday life: Cues, channels, and processes. *Handbook of Communication and Emotion: Research, Theory, Applications, and Contexts*, 29-48.Porter, R. E., & Samovar, L. A. (1996). Cultural influences on emotional

expression: implications for intercultural communication. In *Handbook of Communication and Emotion* (pp. 451-472).

Reed, M. L., & Edelbrock, C. (1983). Reliability and validity of the direct observation form of the child behavior checklist. *Journal of Abnormal Child Psychology, 11*(4), 521-530.

Reschly, A. L., & Christenson, S. L. (2012). Jingle, jangle, and conceptual haziness: Evolution and future directions of the engagement construct. In *Handbook of research on student engagement* (pp. 3-19). Springer, Boston, MA.

Reyes, L., & Fennema, E. (1981). *Classroom Processes Observer Manual*.

Reynolds, C., & Kamphaus, R. (2004). *Behavior Assessment System for Children, (BASC-2) Handout*. AGS Publishing, 4201, 55014-1796.

Richards, L. C., Heathfield, L. T., & Jenson, W. R. (2010). A classwide peer-modeling intervention package to increase on-task behavior. *Psychology in the Schools*, *47*(6), 551-566.

Riley-Tillman, T. C., Kalberer, S. M., & Chafouleas, S. M. (2005). Selecting the right tool for the job: A review of behavior monitoring tools used to assess student response to intervention. *The California School Psychologist, 10*(1), 81-91.

Rodrigo, M.M.T., Baker, R.S.J.d. (2011). Comparing Learners' Affect While Using an Intelligent Tutor and an Educational Game. *Research and Practice in Technology Enhanced Learning, 6* (1), 43-66.

Rodrigo, M. M. T., Baker, R. S., Agapito, J., Nabos, J., Repalam, M. C., Reyes, S. S., & San Pedro, M. O. C. (2012). The effects of an interactive software agent on student affective dynamics while using; an intelligent tutoring system. *IEEE Transactions on Affective Computing, 3*(2), 224-236.

Rodrigo, M.M.T., Baker, R.S.J.d., Rossi, L. (2013) Student Off-Task Behavior in Computer-Based Learning in the Philippines: Comparison to Prior Research in the USA. *Teachers College Record*, 115 (10), 1-27.

San Pedro, M. O. C. Z., d Baker, R. S., & Rodrigo, M. M. T. (2011). Detecting carelessness through contextual estimation of slip probabilities among students using an intelligent tutor for mathematics. In *International Conference on Artificial Intelligence in Education* (pp. 304-311). Springer, Berlin, Heidelberg.

Sayette, M. A., Cohn, J. F., Wertz, J. M., Perrott, M. A., & Parrott, D. J. (2001). A psychometric evaluation of the facial action coding system for assessing spontaneous expression. *Journal of Nonverbal Behavior, 25*(3), 167-185.Sawada, D., Piburn, M. D., Judson, E., Turley, J., Falconer, K., Benford, R., & Bloom, I. (2002).Measuring reform practices in science and mathematics classrooms: The reformed teaching observation protocol. *School Science and Mathematics, 102*(6), 245-253.

Schutz, P. & Pekrun, R. (Eds.) (2007). *Emotion in Education*, Academic Press, New York.

Shapiro, E. S. (1996). *Academic skills problems workbook.* Guilford Press.

Shapiro, E. S., & Heick, P. (2004). School psychologist assessment practices in the evaluation of students referred for social/behavioral/emotional problems. *Psychology in the Schools*, 41, 551-561.

Shapiro, E. S. (2011). Behavior observations of students in schools. In E. S. Shapiro (Ed.), *Academic Skills Problems Fourth Edition Workbook* (pp. 35–56). New York: Guilford Press.

Skinner, C. H., Dittmer, K. I., & Howell, L. A. (2000). *Direct Observation in School Settings: Theoretical Issues.*

Skinner, C. H., Rhymer, K. N., & McDaniel, E. C. (2000). Naturalistic direct observation in educational settings. *Conducting School-Based Assessments of Child and Adolescent Behavior*, 21-54.

Soukup, J. H., Wehmeyer, M. L., Bashinski, S. M., & Bovaird, J. A. (2007). Classroom variables and access to the general curriculum for students with disabilities. *Exceptional Children, 74*(1), 101-120.

Smith, M. K., Jones, F. H., Gilbert, S. L., & Wieman, C. E. (2013). The classroom observation protocol for undergraduate STEM (COPUS): A new instrument to characterize university STEM classroom practices. *CBE-Life Sciences Education, 12*(4), 618-627.

Stanley, S. O., & Greenwood, C. R. (1981). *CISSAR: Code for Instructional Structure and Student Academic Response: Observer's Manual*. Kansas City, University of Kansas, Bureau of Child Research, Juniper Gardens Children's Project.

Steiner, N. J., Sidhu, T., Rene, K., Tomasetti, K., Frenette, E., & Brennan, R. T. (2013). Development and testing of a direct observation code training protocol for elementary aged students with attention deficit/hyperactivity disorder. *Educational Assessment, Evaluation and Accountability, 25*(4), 281-302.

Steiner, N. J., Sidhu, T., Rene, K., Tomasetti, K., Frenette, E., & Brennan, R. T. (2013). Development and testing of a direct observation code training protocol for elementary aged students with attention deficit/hyperactivity disorder. *Educational Assessment, Evaluation and Accountability, 25*(4), 281-302.Tabanao, E., & Rodrigo, M. M. (2016). A comparison of the experience of confusion among Filipino and American learners while using an educational game for physics. *Proceedings of the 24th International Conference on Computers in Education*. India: Asia-Pacific Society for Computers in EducationTsai,

J. L., & Levenson, R. W. (1997). Cultural influences on emotional responding: ChineseAmerican and European American dating couples during interpersonal conflict. *Journal of Cross-Cultural Psychology*, *28*(5), 600-625.

Volpe, R. J., DiPerna, J. C., Hintze, J. M., & Shapiro, E. S. (2005). Observing students in classroom settings: A review of seven coding schemes. *School Psychology Review, 34*(4), 454.Volpe, R. J., & McConaughy, S. H. (2005). Systematic direct observational assessment ofstudent behavior: Its use and interpretation in multiple settings: An introduction to the miniseries. *School Psychology Review, 34*(4), 451.

Walker, H. M., & Severson, H. H. (1990). *Systematic Screening for Behavior Disorders: Users guide and administration manual.* Longmont, CO: Sopris West.

Walkington, C., & Marder, M. (2013). Classroom observation and value added models give complementary information about quality of mathematics teaching. *Designing Teacher Evaluation Systems: New Guidance from the Measures of Effective Teaching Project: San Francisco, Josey Bass*, 234-277.

Waxman, HC, Padron, Y. N. (1994). Alternative Models for Evaluating Technology-Enriched Professional Development Schools. In *Society for Information Technology & Teacher Education International Conference* (pp. 199-202). Association for the Advancement of Computing in Education (AACE).

Waxman, H. C., & Padron, Y. N. (2004). The uses of the Classroom Observation Schedule to improve classroom instruction. *Observational Research in US classrooms: New Approaches For Understanding Cultural and Linguistic Diversity*, 72-96.

Wessel, D. (2015) The potential of computer-assisted direct observation apps. *Int J Interact Mobile Tech, 9*(1), 31.

Wigfield, A., & Cambria, J. (2010). Students' achievement values, goal orientations, and interest: Definitions, development, and relations to achievement outcomes. *Developmental Review*, 30, 1-35.Wilson, M. S., & Reschly, D. J. (1996). Assessment in school psychology training and practice. *School Psychology Review,* 25, 9-23.

Wragg, T. (2002). *An introduction to classroom observation*. Routledge.