Understanding the Impact of Observer Effects on Student Affect

Xiner Liu[1[0009-0004-3796-2251]], Ashish Gurung[2[0000-0001-7003-1476]], Ryan S. Baker[1[0000-0002-3051-3232]], and Amanda Barany[1[0000-0003-2239-2271]]

[1] University of Pennsylvania, Philadelphia, PA 19104, USA
xiner@upenn.edu
[2] Carnegie Mellon University, Pittsburgh, PA 15213, USA

Abstract. The measurement of affect presents a challenge for researchers in quantitative ethnography and related communities, as each of the several possible methods for obtaining ground truth have downsides and often disagree with each other. One common method is field observations; some accounts have raised concerns about observer effects, but further research is needed to understand how much observer effect is present and exactly how observer effects manifest. In this study, we attempt to quantify and detail observer effects in classroom data collected using the Baker Rodrigo Ocumpaugh Monitoring Protocol (BROMP). We leveraged prediction models developed in prior work to assess students' affect and disengaged behavior, during periods where observers were present or not present. Statistical analyses revealed differences in both affective states and behaviors on days when field observations were conducted. When observers were present, students showed increases in concentration and decreases in frustration, off-task behavior, and gaming the system. Findings suggest that the collection of observation data changes the proportion of behavior and affect observed, which should be taken into account when designing, implementing, and analyzing observation-based research. We discuss implications of this finding for quantitative ethnography, when observation data is systematized, visualized, and compared. Our findings do not suggest that the individual observations of specific affective states or behaviors are invalid but offer insights for QE scholars to consider when collecting and analyzing observation data.

Keywords: Observer Effect, Quantifying Observer Effect, BROMP, Affect, Classroom Observation, Field Observation

1 Introduction

Quantitative Ethnographic (QE) research is defined by the formal and systematic examination of data in context to maximize understanding of complex processes while minimizing bias [53]. As such, QE research exploring concepts such as affect (the experience of emotion; emotion in context) and disengagement has continued to expand [34]. Scholars studying affect in the QE community and neighboring communities have employed a variety of techniques to structure or "quantify" data after collection by segmenting it into lines or segments that have interpretable meaning, coding for the presence of that meaning, and grouping coded lines for the purpose of visualization or comparison. Two key issues that impact the interpretability of these results must be addressed, however. First, datasets must be "thick" enough in detail and contextual information for results to reliably represent subtle phenomena such as affect or disengagement. This poses a particular challenge in some cases, as data sources are often incomplete secondhand records of an experience (e.g., transcripts), or are perspectival (e.g. self-report data). Second, the process of systematizing the data positions researchers as the primary "instrument" of knowledge creation [5], which increases the likelihood of issues around fairness and validity, particularly when interpreting secondhand and perspectival data. While practices exist to help address issues of fairness and validity that emerge during these post hoc processes (e.g., inter-rater reliability metrics), they too have been critiqued as inconsistent at best and prone to error at worst [24, 54].

One possible solution to these issues is to conduct field observations [18], in which trained researchers can view and systematically annotate affect in real time. This approach has potential to maximize the contextual information available to the interpreting researcher and helps to ensure that coding and systematization are embedded in the context being studied. Once an ethnographer enters the research context, however, the potential for observer effects (in this case, shifts in the way student affect manifests as a result of being observed) emerges [30]. Observer effects have been recognized as potentially impactful on QE research [53], but further exploration is needed to determine the extent of observer effects and the ways in which they manifest, so that we better understand the impacts they have on the complex behavior and latent constructs such as affect that we are studying. In this study, we attempt to quantify and describe the observer effects that emerged in classroom observation data collected using the Baker Rodrigo Ocumpaugh Monitoring Protocol (BROMP), widely used in other research communities. We use a different method, automated detectors of affect (themselves originally developed using BROMP, but with separate data), to compare students' affect and disengaged behavior during periods when observers were present and when they were not. Our research question is: What are the impacts of observer effects in classroom observation data on how affect and disengagement manifest?

2 Observation Research in QE

The underpinning of QE is high quality data [53]. While the majority of QE research has involved textual data (e.g., [5, 34]) other research in our community has relied upon computer interaction log data [37], observational data [2], and sensor data [4].

However, all data collection introduces distortion [28] as the very act of collecting data impacts the phenomena being studied. For example, placing a camera in a public location can change the behavior of individuals in that location [57], interviews can provoke changes in the participant's thinking [15, 41], and self-report in questionnaires can influence an individuals' emotions (cf. in [9]) and physiological reactions [35].

There has been considerable recent interest in phenomena surrounding affect and disengagement during learning [20, 21, 38], including within the QE community [3, 36]. Research on these constructs predominantly relies on two sources for acquiring coded data for analysis: self-report data (e.g., [32]) and external observations (e.g., [23]). Self-report data offers the advantage of obtaining direct insights from students themselves regarding their learning experiences. However, it is prone to response biases stemming from limited self-awareness [58] or social desirability/demand [40], potentially compromising the accuracy and reliability of the collected information.

External expert judgments can address these limitations, offering more standardized evaluations of students' affective states and behaviors [55]. There are several types of external judgment, each conferring unique benefits and limitations. For instance, text replays, which consist of segments of actions in log data that are reviewed and coded for specific behaviors, have been found to be 2-6 times faster than other labeling methods [8]; but may overlook crucial non-verbal cues such as facial expressions and body movements. By contrast, video replays enable more comprehensive reviews of students' actions and interactions, allowing for convenient rewatches at any given time [51]. However, the processing and coding of substantial video data can be time-consuming and resource-intensive, and the presence of cameras can change behavior [22, 33].

Field observations afford observers an even more in-depth view of students' experiences during learning. During field observations, trained observers, following a specific annotation scheme (cf. [51]), physically attend classroom sessions and systematically record students' affective states and behaviors in real-time. This approach provides observers highly authentic information about students' learning [18], but the presence of observers in the classrooms may lead to observer effects, where students' affective states and behaviors might be influenced due to their awareness of being observed [30].

However, the impact of observer effects has been much more widely reported in an informal fashion than explicitly studied [56] although some exceptions exist. The studies that have attempted to identify explicit quantitative observer effects have in many cases failed to obtain any observer effect at all. For example, Crofoot et al. use radio telemetry to argue that monkey behavior is not different during observations, once monkeys have habituated [19]. Hagel et al. also found that human observers did not change hand-washing behavior compared to behavior measured through logs in automated dispensers [31]. On the other hand, notifying individuals that their social media posts are being monitored appears to substantially change their posting behavior [52]. In addition, the research attempting to quantify observer effects seems to largely focus on straightforward behaviors rather than the more complex behaviors or latent constructs (such as emotions, affect, self-regulation) typically investigated in the QE community.

In this paper, we therefore empirically investigate whether observer effects occur for field observations conducted to study complex behavior and latent constructs. We do so in the context of BROMP (the Baker Rodrigo Ocumpaugh Monitoring Protocol; [12]), one of the most widely used approaches for classroom observation of complex behavior and emotion. In this paper, we investigate the observer effect associated with BROMP through an analysis of interaction log files sourced from the widely used ASSISTments learning platform. We compare the same students both when they were being observed during BROMP and when they were not being observed with BROMP (both earlier and later in the school year). We use previously validated prediction models (themselves initially developed using BROMP) to assess students' affect and disengaged behaviors during these periods – the same constructs being assessed by BROMP, to see if those constructs are impacted by observer effects. Then, we conducted statistical tests comparing the same group of students between the observed and non-observed periods to assess whether the presence of observers influenced the prevalence of emotional or behavioral displays among students. We also investigate whether these differences are consistent across students in urban, rural, and suburban settings.

By developing methods for understanding the impact of observer effects, we can determine the degree to which an observational method produces data which is a reliable indicator of the authentic (unobserved) manifestation of the constructs being studied, and fair to the lived experiences of research subjects [53, 54].

3 The BROMP Protocol

BROMP is a widely used protocol for conducting observations of student affect and complex behavior in educational settings [12]. BROMP utilizes a momentary time sampling approach to code students' behavior and affect, one student at a time, in a predetermined order, in order to ensure a representative sample and avoid focusing on extreme events. During observations, the observers focus on one student until visible affect or behavior is identified or 20 seconds have elapsed, after which they proceed to the next student. The observers make a comprehensive and holistic evaluation of the students' affect and behavior, considering their facial expression, posture, utterances, activity, and other factors. Observations are typically recorded using a custom-built application called HART (Human Affect Recording Tool; [46]), which provides precise time stamps of each observation to facilitate unobtrusive data collection.

The influence of BROMP is not limited to the United States alone [12]. With over 150 certified coders in 7 countries, BROMP has been applied to study a wide range of phenomena, such as learners' cognition [43], on-task reflection [29], teachers' proactive remediation [42], affect sequences [39], and the effect of engagement interventions [7]. BROMP has been used in several studies to generate training labels for the development of automated detectors, including detectors of affective states [16] and disengaged behaviors such as gaming the system [10], off-task behavior [17], and wheel-spinning [47]. These detectors, in turn, have been used for a number of purposes, including to inform and contextualize interviews [3, 11] and also to study phenomena such as participation in stem-related careers [1]. Overall, BROMP has generated data which has been extensively used in many research communities, including QE [3, 39].

BROMP's design integrates several strategies to mitigate the observer effect [12]. Before sessions, observers coordinate with teachers to minimize extra interaction and clarify their role to students if needed. During observations, observers maintain a neutral and unobtrusive disposition to minimize disruptions, including avoiding eye contact, not looking directly at the student being observed, dressing in bland colors, and moving slowly ("mosey don't walk"). If a student becomes aware of being observed, the data for that particular observation is discarded to ensure the integrity of the data. Furthermore, the HART app allows labeling with small handheld devices, which are relatively unobtrusive and hard for students to see. However, it is not clear whether these measures have been successful at addressing observer effects.

4 Context

In this study, we investigated the potential observer effect of conducting BROMP observations within the ASSISTments platform, an online learning system that is primarily focused on mathematics. With a user base of around 500,000 students and 20,000 teachers, ASSISTments is widely used, primarily in the United States but also in over 20 other countries [26]. The platform breaks down problems into steps, provides hints at each stage, and offers the answer through a "bottom-out hint" when requested. The assessment process offers assistance, scaffolding, and feedback to students, while equipping teachers with detailed information about students' knowledge and performance for targeted assistance. Several randomized controlled studies have shown significant learning gains for students who regularly engage with the platform [25, 27, 44].

4.1 Dataset and Population

ASSISTments provides substantial support for external researchers by offering publicly available datasets that include rich interaction log data, along with field observation data and longitudinal student outcomes. To investigate the potential impact of classroom observers on students' affective experiences and behaviors during online learning, we used both the interaction log data and the data on field observations conducted during the same years, obtaining the field observation data from the ASSISTments webpage (https://sites.google.com/site/assistmentsdata/) and other data from the same year directly from the ASSISTments team, by request. The field observation data were systematically collected by human observers during in-class computer lab sessions [45]. During observations, the handheld devices for affect labeling were synchronized to the educational software logging server. Leveraging the time stamp and user ID in the interaction log file, we were able to select students who were part of the field observations and discern whether specific actions by the student occurred within or outside the days when observation occurred.

Our secondary analysis involves a sample of middle school students from three distinct populations in the Northeastern region of the United States, with data originally reported in a prior study by [45]. The first group, drawn from two schools in Maine, consists of predominantly white students from a rural background with low socioeconomic status, over 50% of whom receive free or reduced-price lunch (which is commonly used as an indicator of poverty in the United States). The second group, drawn from three suburban schools in Massachusetts, consists mainly of mid-to-high socioeconomic-status White and East-Asian students, with less than 20% receiving free or reduced-price lunches. The third group of students came from an urban setting in Massachusetts and comprises primarily lower-income Latinos/students of Puerto Rican origin, with English as their first language, as well as African American and Balkan students, over half of whom received free or reduced-price lunch. Inclusion of these distinct populations in the study sample enables us to comprehensively investigate potential differential impacts of human observers on various student populations.

5 Method

5.1 Inferring Students' Affective States

The primary goal of this study was to quantify observer effects and study their magnitude during classroom observations, focusing on potential changes in the frequency of students' affective states and disengaged behaviors. To accomplish this goal, we derived students' affective experiences and behavioral responses during both the observation and non-observation periods; the same constructs being studied using BROMP.

We inferred students' affective states (i.e., boredom, engaged concentration, frustration, and confusion) using the interaction log files data from ASSISTments. The interaction log files were pre-processed and run through the ASSISTments affect detectors from [16], which utilized the Long-Short Term Memory (LSTM) algorithm. Evaluated through a rigorous 5-fold cross-validation procedure at the student level, these detectors achieved an average AUC ROC of 0.77 across four affective states (AUC ROC is the most relevant metric for our current analyses, since our analyses aggregate across detector confidences), as shown in Table 1. It is worth noting that these detectors were originally developed using BROMP observations (on other data) as ground truth. This choice enables us to avoid confounds if ground truth obtained in other fashions (such as self-report) lead to the constructs of interest being defined subtly differently.

The source code of these models was made available to us by the ASSISTments team. This code computes the confidence scores for each affective state at the level of 20-second clips. For example, if the engaged concentration detector gives an output of 0.67, it signifies a moderate level of confidence, indicating that the detector believes that there is a 67% probability that the student is experiencing engaged concentration based on their actions. To ensure comparability among the affect detectors, we implemented a normalization procedure. This involved subtracting the minimum value from each score and dividing it by the difference between the maximum and minimum confidence values for that state (across all data). Within each clip, we then averaged across all clips in a problem to identify the most prevalent affective state for each problem.

5.2 Inferring Students' Disengaged Behaviors

We used the same interaction log data for identifying students' disengaged behaviors, consisting of instances of being off-task and gaming the system. Although machine-learned detectors of gaming the system and off-task behavior have been developed for ASSISTments [50], these detectors are not currently available from the ASSISTments platform. Therefore, we used different models of gaming the system (e.g. [49]) and off-task behavior [6]. These predictions were also generated at the clip level. However, a different clip size was used for these detectors, based on the way the gaming detector in [49] was developed. In ASSISTments, when students are presented with an "original" problem, they are expected to provide the answer without needing to detail individual steps, as long as they solve the problem correctly on their initial attempt. However, if students fail to provide the correct answer, they may need to answer "scaffolding questions" correctly in order to successfully complete the problem. As a result, a clip was defined as the sequence of actions starting from the first action on an original unscaffolded problem to the last attempt before the subsequent original unscaffolded problem. This definition implies that a clip can consist of just one action or more than 50 actions. [49] developed a cognitive model leveraging knowledge engineering to discern instances of gaming behavior. In our study, we applied this detector to the current data set to obtain predictions at the same clip level as previously defined.

For off-task behavior, we use a model proposed by [6] and also used as a feature in the machine-learned ASSISTments model [50]. [6] compared a machine-learned detector of off-task behavior to a simpler approach that used a fixed time cutoff of 80 seconds. While the sophisticated model outperformed the cutoff model in distinguishing between on-task and off-task behavior, the cutoff model still exhibited a reasonable correlation of 0.46 with classroom observations of off-task behavior. Given the unavailability of the machine learned model for our current data set, we used this cutoff model to identify students' off-task behavior, classifying clips as involving off-task behavior if any action within the clip exceeded 80 seconds.

5.3 Data Aggregation

The data were filtered to include clips from active student users who were present both before, during, and after the observations. We then created a time period variable based on the timestamps, which had a value of 1 if the students were using the system during the field observations (same day) and 0 otherwise. Using these filtered datasets, we constructed two pivot tables to calculate the clip-level percentages of learners' affective states or behaviors for both periods. For affective states, the dataset was grouped based on user ID, and periods. Then, four separate columns were created to represent different affect categories: boredom, concentration, frustration, and confusion. The same approach was employed for summarizing off-task and gaming behaviors. Within each period, we calculated the occurrence count for each affective state/behavior and expressed it as a percentage relative to the total number of clips.

5.4 Statistical Analysis and Benjamini-Hochberg Procedure

Prior to conducting the statistical test, we examined the normality assumption of affect/behavior percentages through visual inspections of histograms and normal probability plots. The outcomes indicated notable deviations from a normal distribution. Concentration appeared to be highly left-skewed, while boredom, frustration, and confusion showed highly right-skewed distributions. These findings are consistent with the observations reported by [16]. Off-task percentages displayed a relatively uniform distribution ranging from 0 to 1 for both observed and non-observed periods. The percentages of gaming the system behaviors, on the other hand, displayed right-skewness for both periods, aligning with the findings of [49].

To further validate the results, we performed the Shapiro-Wilk test to assess the normality assumptions; results reaffirmed that the percentages do not follow a normal distribution ($p < 0.01$). Therefore, given the structure of the data and the presence of paired students across the observed and non-observed periods, we proceeded with using the Wilcoxon Matched-Pairs Signed-Ranks Test, which is suitable for non-parametric data analysis and dependent samples, without relying on the assumption of normality.

In assessing our findings, we employed the Benjamini-Hochberg procedure (B&H; [14]) which is designed to control the false discovery rate (FDR) when multiple tests are conducted. Benjamini-Hochberg was used instead of Benjamini-Yekutieli due to the lack of positive regression dependency between tests. When applying B&H, we identified statistically significant findings while maintaining control over the overall occurrence of false positives. The procedure involves arranging the obtained p-values in ascending order and comparing them to critical thresholds that are adjusted based on the number of tests conducted (i.e., adjusted alpha value). These critical thresholds guarantee that the expected proportion of false discoveries among all rejected hypotheses remains below the predetermined FDR level (0.05). Specifically, the first p-value is compared to a threshold of (0.05 divided by the number of tests), the second p-value is compared to (0.05 divided by the number of tests minus 1), and so forth, until the last p-value is compared to 0.05. If a p-value is less than or equal to its corresponding critical threshold, it indicates statistical significance at the chosen FDR level. Alternatively, a p-value greater than the critical threshold but less than twice that threshold indicated marginal significance at the chosen level. In our study, we investigated the potential observer effect on four affective states and two learning behaviors, leading to 6 statistical tests across all regions and 18 statistical tests across the three subregions. We compared the obtained p-values from tests conducted on all regions or subregions to their adjusted alpha values based on 6 and 18 tests, respectively.

6 Results

6.1 Observer Effect on Learners' Affective States

Table 2 presents the statistical results for affect percentages during and not during observations. We first consider boredom. Overall, no significant differences in boredom percentages were observed among all students considered together. When analyzing students from specific subregions, boredom levels did not significantly differ for students from suburban and urban areas. Students in rural regions appeared to be less bored during the observations than outside of them (p = 0.036), but this difference becomes only marginally significant after applying the B&H post-hoc correction (α = 0.035). By contrast, the results revealed an increase in concentration percentages during observations when all regions are considered together (p = 0.049), which remained marginally significant after the B&H correction ($\alpha_{adjusted}$ = 0.067). The results indicated that the presence of observers in the classroom seemed to be associated with higher levels of

Table 1. Statistical Results for Affect Percentages During and Not During Observations

Affective States	Region	Sample Size	W Statistics	p-value	Median Not During	Median During
Boredom	All	202	15262	0.792	0.04	0.05
	Rural	10	16	0.036	0.05	0.02
	Suburban	125	5368	0.781	0.03	0.04
	Urban	67	1841	0.626	0.07	0.08
Engaged	All	202	16258	0.049*	0.83	0.86
	Rural	10	34	0.151	0.88	0.90

Concentration	Suburban	125	6409	0.475	0.81	0.87
	Urban	67	1782	0.029	0.85	0.86
Frustration	All	202	766	0.019**	0.00	0.00
	Rural	14	7	0.462	0.00	0.00
	Suburban	155	119	0.004**	0.00	0.00
	Urban	79	199	0.927	0.00	0.00
Confusion	All	202	12782	0.097	0.06	0.06
	Rural	10	33	0.363	0.04	0.06
	Suburban	125	5798	0.866	0.06	0.07
	Urban	67	918	0.007**	0.06	0.05

*Marginally Significant after B&H correction
**Significant after B&H correction

student concentration. However, there were no significant differences in concentration levels between periods within rural and suburban regions, and the apparent decrease in confidence observed in urban regions did not retain significance after the correction (p = 0.029, α = 0.014).

A significant difference was also found in frustration levels between observed and non-observed periods when all regions were considered together (p = 0.019). Specifically, distribution plots (Figure 1) revealed that students experienced lower levels of frustration during observations compared to when they were not being observed. In specific, the mean during the observation period was 0.004, whereas the mean outside the observation period was 0.008. When examining the results for individual regions, only the suburban region showed a significant difference in frustration levels (p = 0.004). For suburban learners, the mean during the observation period was 0.003, whereas the mean outside the observation period was 0.007. This finding is shown in the distribution plot on the right side of Figure 1.

No significant differences were observed in confusion levels between observed and non-observed periods when all regions were considered together. This was also seen in rural and suburban schools. However, a significant decrease in confusion level was observed during the observation period within students in urban regions (p = 0.007).

6.2 Observer Effect on Learners' Behaviors

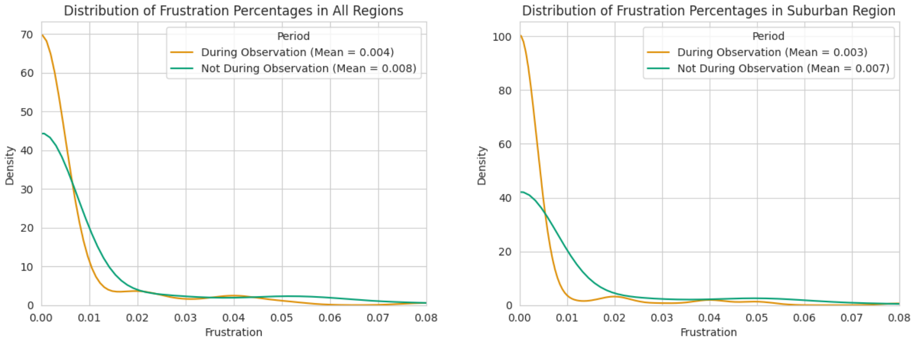
Distribution of Frustration Percentages in All Regions
Distribution of Frustration Percentages in Suburban Region

Fig. 1. Distribution of Frustration Percentages in All regions (*left*) and Suburban Region (*right*)

We next present the statistical results for disengaged behavior percentages during and not during observations in Table 3. The analysis of off-task behavior revealed significantly less off-task behavior for all regions were considered together when observers were present in the classroom (p < 0.001), a finding also seen in the urban region (p = 0.005) but not in rural and suburban regions.

Table 2. Statistical Results for Behavior Percentages During and Not During Observations

Affective States	Region	Sample Size	W Statistics	p-value	Median Not During	Median During
Off-task	All	717	106151	0.000**	0.299	0.270
	Rural	100	1885	0.143	0.276	0.231
	Suburban	288	17832	0.083	0.227	0.220
	Urban	329	22336	0.005**	0.400	0.333
Gaming the System	All	717	17029	0.000**	0.000	0.000
	Rural	100	275	0.000**	0.005	0.000
	Suburban	288	925	0.249	0.000	0.000
	Urban	329	6831	0.047	0.000	0.000

*Marginally Significant after B&H correction
**Significant after B&H correction

A significant though small decrease in gaming the system behavior occurred during observations when all regions were considered together (p < 0.001). The mean amount of gaming during the observation period was 0.02, while the mean amount of gaming outside the observation period was 0.03. This decrease was also seen in rural areas (p < 0.001), while no significant difference was observed for urban and suburban schools.

7 Discussion & Conclusion

Within quantitative ethnography, our methods draw conclusions that are shaped by the codes that we obtain and the structure we set. The quality of those choices has a significant impact on the validity of the conclusions we can draw, and whether they are fair to phenomena and individuals being studied [54]. All active measurements impact the subjects being studied in some way, but the impact of measurements on the phenomena being studied are themselves under-studied. We attempted to quantify the degree to which classroom observations conducting using the Baker Rodrigo Ocumpaugh Monitoring Protocol (BROMP) have observer effects, and how the observer effect impacts the prevalence of different affective experiences and behaviors observed among students. Our findings revealed significant observer effects: a noticeable increase in engaged concentration, as well as a decrease in frustration, off-task behavior, and gaming the system when human observers were present in the classroom. Some differences in observer effects were found between demographic cohorts of students – better understanding these differences might require interviews of students during or after observations.

The existence of an observer effect raises important questions about the meaning and validity of codes collected through BROMP and similar methods. One possible interpretation is that the presence of human observers has actually changed the distribution of affective states and behavior among students. For example, the presence of more adults in the room might prompt students to focus more on the learning task, producing all the effects seen. In other words, students might genuinely be gaming the system less and taking the learning task more seriously. If this is the case, then BROMP observation is not changing how behaviors and affect manifest but is instead changing how much (and perhaps when) behaviors and affect occur. This would not substantially impact the validity of automated detectors trained using these codes but might impact the validity of conclusions drawn from analyzing BROMP data directly in QE research.

If the observer effect only alters the proportion of observed affect and behavior in a dataset, the impact on visual and statistical analyses of the data is likely to produce consistent findings, particularly for epistemic networks, which are normalized to account for different unit sizes. However, if observer effects instead lead to changes in how the same affect and behavior manifest, it could lead to more serious consequences for reliability and validity. For example, our results could also be due to students consciously concealing or feigning different affect than what they are genuinely experiencing. In this case, even though the students may appear to be more focused or less frustrated during observation, their expressions may stem solely from self-presentation considerations. Consequently, merely observing the students may fail to provide an accurate reflection of their true emotional and behavioral dynamics, as their external reactions no longer authentically represent their inner experiences. The labels collected through these observations would not truly represent the learners' states and conditions. If this were the case, it becomes relevant to ask what the previous correlations between BROMP-derived detectors and other measures means. Is engaged concentration genuinely associated with better learning (e.g., [50])? Or is a combination of genuine engaged concentration and the situational awareness to fake it what is actually predictive?

While our study provides evidence for an observer effect, it is unclear whether BROMP observation alters the proportion of affect and behavior or whether BROMP is modifying their externally visible manifestations. Still, if BROMP is being used to compare the proportion of different affect or disengaged behaviors between systems or contexts, it can only safely be compared to data also collected using BROMP or very similar methods. This recommendation aligns with QE analysis best practices, where only datasets from similar sources and of similar structure can be reliably compared.

However, our findings do not yet indicate that there is risk involved in using BROMP to develop detectors of affect, as we have yet not seen evidence that BROMP changes how behaviors and affect manifest in log data. To investigate that foundational question, further research and triangulation will be necessary. This may involve finding a method with highly similar operationalizations but without such an observer effect. For example, retrospective hand-labeling of interaction data has been extensively used to identify some disengaged behaviors (e.g. [48]). At the moment, we are not aware of an approach for affect coding that lacks such limitations, as self-report produces demand effects, video observation also produces observer effects [13], and observing students through video without their awareness would generally be considered unethical.

In addition, while data collected using BROMP showed evidence of observer effects, there is no evidence that observer effects using BROMP are higher than the distortions that might be produced by competing methods, such as video, other classroom observation methods, and self-report. For all we know, BROMP may be less impacted than other approaches; determining this requires evaluating the impact of alternative label collection methods on the same types of constructs within the same types of settings. The comparison would benefit the QE field by determining how significant the observer effects are for a range of data collection methods that can be used by the community.

However, while the precise nature and extent of the observer effect in BROMP and other methods remains uncertain, our findings indicate that greater attention to this phenomenon is necessary. While it is improbable that we can entirely eliminate observer effects, greater consideration of this challenge may enable us to better understand the limitations of our current data and findings, and over time, may help us to increase the validity of our methods and the quality of our data. In turn, this will help to guarantee that quantitative ethnography analyses meet their goals for fairness and validity.

Acknowledgement

We express our sincere appreciation to the ASSISTments team for providing us with the interaction log data, BROMP data, and source code for the affect detectors. Thank you to Dr. Luc Paquette for sharing the source code of the gaming system detector transferred to the ASSISTments platform and for assisting with its application. We are grateful to Stefan Slater for conducting a thorough code review for this project.

References

1. Almeda, M. V., & Baker, R. S.: Predicting student participation in STEM careers: The role of affect and engagement during middle school. Journal of Educational Data Mining, 12(2), pp. 33-47 (2020)
2. Andres-Bray, T., Barany, A., & Gonder, M. K.: Using epistemic network analysis to explore flexibility and development of termite fishing techniques in nigeria-cameroon chimpanzees (pan troglodytes ellioti). Proc. of the Int'l Conf. on Quantitative Ethnography (2023)
3. Andres, J.M.A.L., Hutt, S., Ocumpaugh, J., Baker, R.S., Nasiar, N., Porter, C.: How Anxiety affects affect: A quantitative ethnographic investigation using affect detectors and data-targeted interviews. Proc. of the Int'l Conf. on Quantitative Ethnography (2021)
4. Andrist, S., Collier, W., Gleicher, M., Mutlu, B., & Shaffer, D.: Look together: Analyzing gaze coordination with epistemic network analysis. Frontiers in psychology, 6, 144911 (2015)
5. Arastoopour Irgens, G., & Eagan, B.: The foundations and fundamentals of quantitative ethnography. Proc. of the Int'l Conf. on Quantitative Ethnography, pp. 3-16 (2022)
6. Baker, R. S.: Modeling and understanding students' off-task behavior in intelligent tutoring systems. Proc. ACM SIGCHI Conf., pp. 1059-1068 (2007)
7. Baker, R. S., & Beck, J.: Adapting to when students game an intelligent tutoring system. Proc. of the 8th Int'l Conf. on Intelligent Tutoring Systems, pp. 392-401 (2006)
8. Baker, R. S., Corbett, A. T., & Wagner, A. Z.: Human classification of low-fidelity replays of student actions. Proc of the Educational Data Mining Workshop at the 8th Int'l Conf. on Intelligent Tutoring Systems, Vol. 2002, pp. 29-36 (2006)
9. Baker, R. S., D'Mello, S.K., Rodrigo, M.M.T., & Graesser, A.C.: 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. Int'l Journal of Human-Computer Studies, 68 (4), pp. 223-241 (2010)

10. Baker, R.S., Corbett, A.T., & Koedinger, K.R.: Detecting student misuse of intelligent tutoring systems. Proc. of the 7th Int'l Conf. on Intelligent Tutoring Systems, pp. 531-540 (2004)
11. Baker, R.S., Hutt, S., Bosch, N., Ocumpaugh, J., Biswas, G., Paquette, L., Andres, J.M.A., Nasiar, N., Munshi, A.: Detector-driven classroom interviewing: Focusing qualitative researcher time by selecting cases in situ. To appear in Educational Technology Research & Development (in press)
12. Baker, R.S., Ocumpaugh, J.L., & Andres, J.M.A.L.: BROMP quantitative field observations: A review. Learning Science: Theory, Research, and Practice, pp. 127-156 (2020)
13. Becker, T. E., & Marique, G.: Observer effects without demand characteristics: An inductive investigation of video monitoring and performance. Journal of Business and Psychology, 29, pp. 541-553 (2014)
14. Benjamini, Y., & Hochberg, Y.: Controlling the false discovery rate: A practical and powerful approach to multiple testing. Journal of the Royal Statistical Society: Series B (Methodological), 57(1), pp. 289-300 (1995)
15. Bosch, N., Zhang, Y., Paquette, L., Baker, R., Ocumpaugh, J., & Biswas, G.: Students' verbalized metacognition during computerized learning. Proc. of the 2021 ACM SIGCHI Conf. on Human Factors in Computing Systems, pp. 1-12 (2021)
16. Botelho, A.F., Baker, R., & Heffernan, N.: Improving sensor-free affect detection using deep learning. Proc. of 18th Int'l Conf. on Artificial Intelligence in Education, pp. 40-51 (2017)
17. Carpenter, D., Emerson, A., Mott, B.W., Saleh, A., Glazewski, K.D., Hmelo-Silver, C.E., & Lester, J.C.: Detecting off-task behavior from student dialogue in game-based collaborative learning. Proc. 21st Int'l Conf. on Artificial Intelligence in Education, pp. 55-66 (2020)
18. Cleary, T.J., & Platten, P.: Examining the correspondence between self-regulated learning and academic achievement: A case study analysis. Education Research Int'l, pp. 1-18 (2013)
19. Crofoot, M. C., Lambert, T. D., Kays, R., & Wikelski, M. C.: Does watching a monkey change its behaviour? Quantifying observer effects in habituated wild primates using automated radiotelemetry. Animal Behaviour, 80(3), pp. 475-480 (2010)
20. D'Mello, S. K., Moulder, R. G., & Jensen, E.: Momentary measures of emotions during technology-enhanced learning prospectively predict standardized test scores in two large samples. Learning and Instruction, 90, 101872 (2024)
21. de Morais, F., & Jaques, P. A.: Dinâmica de afetos dos alunos em um sistema tutor inteligente de matemática no contexto brasileiro. In Anais do XXXII Simpósio Brasileiro de Informática na Educação, pp. 691-704 (2021)
22. Derry, S. J., Pea, R. D., Barron, B., Engle, R. A., Erickson, F., Goldman, R., ... & Sherin, B. L.: Conducting video research in the learning sciences: Guidance on selection, analysis, technology, and ethics. The journal of the learning sciences, 19(1), pp. 3-53 (2010)
23. Dragon, T., Arroyo, I., Woolf, B. P., Burleson, W., El Kaliouby, R., & Eydgahi, H.: Viewing student affect and learning through classroom observation and physical sensors. Proc. of the 9th Int'l Conf. on Intelligent Tutoring Systems, pp. 29-39 (2008)
24. Eagan, B., Brohinsky, J., Wang, J., & Shaffer, D. W.: Testing the reliability of inter-rater reliability. Proc. of the Int'l Conf. on Learning Analytics & Knowledge (2020)
25. Fairman, J., Porter, M., & Fisher, S.: Principals discuss early implementation of the ASSISTments online homework tutor for mathematics. ASSISTments Efficacy Study Report 2 2015)

26. Feng, M., Heffernan, N., Collins, K., Heffernan, C., & Murphy, R. F.: Implementing and evaluating ASSISTments online math homework support at large scale over two years: Findings and lessons learned. Proc. Int'l Conf. on Artificial Intelligence in Education (2023)
27. Feng, M., Roschelle, J., Heffernan, N., Fairman, J., & Murphy, R.: Implementation of an intelligent tutoring system for online homework support in an efficacy trial. Proc. of the 12th Int'l Conf. on Intelligent Tutoring Systems, pp. 561-566 (2014)
28. Geertz, C.: The interpretation of cultures (Vol. 5019). Basic books (1973)
29. Grawemeyer, B., Mavrikis, M., Holmes, W., Gutiérrez-Santos, S., Wiedmann, M., & Rummel, N.: Affective learning: Improving engagement and enhancing learning with affect-aware feedback. User Modeling and User-Adapted Interaction, 27, pp. 119-158 (2017)
30. Gupta, A., D'Cunha, A., Awasthi, K., & Balasubramanian, V.: Daisee: Towards user engagement recognition in the wild. arXiv preprint arXiv:1609.01885 (2016)
31. Hagel, S., Reischke, J., ... & Pletz, M. W.: Quantifying the hawthorne effect in hand hygiene compliance through comparing direct observation with automated hand hygiene monitoring. Infection Control & Hospital Epidemiology, 36(8), pp. 957-962 (2015)
32. Hutt, S., Grafsgaard, J.F., & D'Mello, S. K.: Time to scale: Generalizable affect detection for tens of thousands of students across an entire school year. Proc. of the 2019 CHI Conf. on Human Factors in Computing Systems, pp. 1-14 (2019)
33. Jansen, A. M., Giebels, E., Van Rompay, T. J., & Junger, M.: The influence of the presentation of camera surveillance on cheating and pro-social behavior. Frontiers in psychology, 9, 302214 (2018)
34. Kaliisa, R., Misiejuk, K., Irgens, G. A., & Misfeldt, M.: Scoping the emerging field of quantitative ethnography: opportunities, challenges and future directions. Proc. of the Int'l Conf. on Quantitative Ethnography, pp. 3-17 (2021)
35. Kassam, K.S., & Mendes, W.B.: The effects of measuring emotion: Physiological reactions to emotional situations depend on whether someone is asking. PloS One, 8(6), e64959 (2013)
36. Karumbaiah, S., & Baker, R. S.: Studying affect dynamics using epistemic networks. Proc. of the Int'l Conf. on Quantitative Ethnography, pp. 362-374 (2021)
37. Karumbaiah, S., Baker, R. S., Barany, A., & Shute, V.: Using epistemic networks with automated codes to understand why players quit levels in a learning game. Proc. of the Int'l Conf. on Quantitative Ethnography, pp. 106-116 (2019)
38. Karumbaiah, S., Baker, R. S., Ocumpaugh, J., & Andres, J. M. A. L.: A re-analysis and synthesis of data on affect dynamics in learning. IEEE Transactions on Affective Computing, 14(2), pp. 1696-1710 (2021)
39. Karumbaiah, S., Baker, R.S.: Studying affect dynamics using epistemic networks. Proc. of the Int'l Conf. on Quantitative Ethnography, pp. 362-374 (2020)
40. Krumpal, I.: Determinants of social desirability bias in sensitive surveys: A literature review. Quality & Quantity, 47(4), pp. 2025-2047 (2013)
41. Kvale, S.: Inter Views: An introduction to qualitative research interviewing. Thousand Oaks, CA: Sage (1996)
42. Miller, W.L., Baker, R., Labrum, M., Petsche, K., Liu, Y-H., & Wagner, A.: Automated detection of proactive remediation by teachers in reasoning mind classrooms. Proc. of the 5th Int'l Learning Analytics and Knowledge Conf., pp. 290-294 (2015)
43. Munshi, A., Rajendran, R., Ocumpaugh, J., Biswas, G., Baker, R. S., & Paquette, L.: Modeling learners' cognitive and affective states to scaffold SRL in open-ended learning environments. Proc. of the 26th Conf. on User Modeling, Adaptation, and Personalization (2018)

44. Murphy, R., Roschelle, J., Feng, M., & Mason, C. A.: Investigating efficacy, moderators and mediators for an online mathematics homework intervention. Journal of Research on Educational Effectiveness, 13(2), pp. 235-270 (2020)
45. Ocumpaugh, J., Baker, R., Gowda, S., Heffernan, N., & Heffernan, C.: Population validity for Educational Data Mining models: A case study in affect detection. British Journal of Educational Technology, 45(3), pp. 487-501 (2014)
46. Ocumpaugh, J., Baker, R.S., Rodrigo, M.M.T., Salvi, A. van Velsen, M., Aghababyan, A., & Martin, T.: HART: The human affect recording tool. Proc. of the ACM Special Interest Group on the Design of Communication (SIGDOC) (2015)
47. Palaoag, T. D., Rodrigo, M. M. T., Andres, J. M. L., Andres, J. M. A. L., & Beck, J. E.: Wheel-spinning in a game-based learning environment for physics. Proc. Int'l Conf. on Intelligent Tutoring Systems, pp. 234-239 (2016)
48. Paquette, L., & Baker, R. S.: Comparing machine learning to knowledge engineering for student behavior modeling: A case study in gaming the system. Interactive Learning Environments, 27(5-6), pp. 585-597 (2019)
49. Paquette, L., Baker, R. S., De Carvalho, A. & Ocumpaugh, J.: Cross-system transfer of machine learned and knowledge engineered models of gaming the system. Proc. of the 22nd Conf. on User Modeling, Adaptation, and Personalization, pp. 183-194 (2015)
50. Pardos, Z.A., Baker, R.S.J.d., San Pedro, M.O.C.Z., Gowda, S.M., & Gowda, S.M.: 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), pp. 107–128 (2014)
51. Porayska-Pomsta, K., Mavrikis, M., D'Mello, S., Conati, C., & Baker, R. S.: Knowledge elicitation methods for affect modelling in education. Int'l Journal of Artificial Intelligence in Education, 22(3), pp. 107-140 (2013)
52. Saha, K., Gupta, P., Mark, G., Kıcıman, E., & De Choudhury, M.: Observer Effect in Social Media Use. Unpublished manuscript, retrieved May 2, 2024 from https://www.researchsquare.com/article/rs-2492994/v3 (2024)
53. Shaffer, D. W.: Quantitative ethnography. Lulu. com (2017)
54. Shaffer, D. W., & Ruis, A. R.: How we code. Proc. of the Int'l Conf. on Quantitative Ethnography, pp. 62-77 (2021)
55. Spencer, C., Koc, I. A., Suga, C., Lee, A., Dhareshwar, A. M., Franzén, E., Lozzo, M., Morrison, G., & McKeown, G.: Assessing the use of physiological signals and facial behaviour to gauge drivers' emotions as a UX metric in automotive user studies. 12th Int'l Conf. on Automotive User Interfaces and Interactive Vehicular Applications, pp. 78-81 (2020)
56. Sporrong, S. K., Kalleberg, B. G., Mathiesen, L., Andersson, Y., Rognan, S. E., & Svensberg, K.: Understanding and addressing the observer effect in observation studies. In Contemporary Research Methods in Pharmacy and Health Services, pp. 261-270 (2022)
57. Van Rompay, T. J., Vonk, D. J., & Fransen, M. L.: The eye of the camera: Effects of security cameras on prosocial behavior. Environment and Behavior, 41(1), pp. 60-74 (2009)
58. Widen, S. C., & Russell, J. A.: Children acquire emotion categories gradually. Cognitive Development, 23(2), pp. 291-312 (2008)