A Case for (Inter)Action: The Role of Log Data in QE

Jennifer Scianna [1] [0000-0003-1029-3452], Xiner Liu [2] [0009-0004-3796-2251], Stefan Slater [2] [0000-0002-1016-1516], and Ryan S. Baker [2] [0000-0002-3051-3232]

[1] University of Wisconsin, Madison WI 53711, USA
[2] University of Pennsylvania, Philadelphia PA 19102, USA
jscianna@wisc.edu

Abstract. Digital tools have the ability to log the fine-grained details of user experiences within and across the system. These digital experiences can lend valuable contextualization to other ethnographic insights. In this paper, we discuss the potential for using interaction logs as a data source and the pipeline considerations that can facilitate and enhance quantitative ethnographic research using this type of data. We draw on previous QE work and examples from QE adjacent fields such as educational data mining, learning analytics, and human-computer interaction to provide evidence for this approach.

Keywords: Interaction Data, Ethno-Mining, Automated Codes

1 Introduction

Personal and handheld computing has increased the ubiquity of technology as an element in many daily interactions. It is not uncommon to see people engaging in person while simultaneously also sharing media through their devices. This new level of engagement indicates that there may be social elements unaccounted for when conducting ethnographic work that does not include the technical elements of the sociotechnical system. In addition, an increasing amount of deep interaction occurs between humans and computers, particularly as people interact in new -- increasingly social and relational -- fashions with technologies [1].

Quantitative ethnography (QE) research has engaged fields where these forms of interaction are commonplace. Medical simulations [2], educational platforms [3–5], entertainment games [6], and social media [7, 8] have all been utilized as both site and object for QE studies. As researchers have moved into these more technology-mediated domains, there has been a call for techniques that can better incorporate the accompanying data streams that are available.

Interaction logs have been used to a greater degree in QE-adjacent communities, such as educational data mining (EDM), learning analytic (LA), and human-computer interaction, for many years. These areas draw from the rich, facilitated user interactions with both the system and other users, to explore the emergence of a highly contextualized digital world that operates in parallel to the “real” embodied and internal worlds of participant and researcher [9]. Fortunately, log files generated from these interactions provide valuable insights into the nature of this world and how users nav-

igate it. Take for example previous QE work exploring user identity development as discussed through discussion forums. This work made inference solely from players' posts about their gameplay, work that could have been augmented by an analysis of the gameplay between posts [6]. Connections to in-game experiences might enhance the analysis by allowing for more grounded exploration of player behavior alongside their meta-reflections. This type of grounded exploration of player behavior in the context of their interaction with a game is seen in [4], which relates player strategies, identified through qualitative coding, to the implicit feedback the game provided.

Beyond the QE community (and prior to its advent among scholars now active in QE), interaction logs have been shown to provide valuable insights into the situational context that impacts user behavior [10], enabling the identification of patterns and anomalies in decision-making processes [11], and offering a window into the user's affective states and reactions to the system or environment [12, 13]. In these projects, the work of interpreting interaction logs can be seen as ethnographic in that it involves analyzing data to uncover insights about how individuals interact within cultures using digital tools. Thus, a thorough examination of interaction log data in relevant contexts may provide more insights for QE research.

The considerable uptake of interaction logs within QE-adjacent communities provides an additional opportunity for interaction log data -- the possibility of extending on what can easily be accomplished through human coding, to scale across much larger corpuses of data. Like the work on extending human coding of text through tools like nCoder [14], automated detectors of user interaction developed through machine learning can enable the analysis of patterns of interest across contexts, using tools like epistemic network analysis [e.g. 3, 4].

In this paper, we first situate our work in relation to other thinking on digital data and ethnography. We then follow with a discussion of the ways in which interaction data has been utilized in QE and how expanded inclusion of this data may further augment prior QE research. We discuss two ways to develop qualitative codes on interaction data: through evidence-centered approaches and extending upon them using machine learning. Ultimately, through comparing each of these approaches to more established paradigms for QE, we discuss why machine learning codes are particularly useful for understanding interaction, and we conclude with a discussion of the compromises inherent to the use of this practice in QE.

2 Digital Data and Ethnography

Shaffer & Hod [15] reiterated the importance of ethnography as the focal point of QE research during the 2022 conference in Copenhagen, stating that the tools and methods being used were directly a response to the need for ethnographers to be able to capture the interactions of culture. As part of exploring the many ways interaction logs may be utilized for QE research, it may be valuable to first step back and consider the role that digital data has had in ethnography itself.

Haines [9] identifies a trajectory of ethnographic work that begins with research that considers digital as the subject of inquiry, which involves studying how digital

technologies shape people's behaviors and experiences, such as through netnography (online ethnography of digital communities) [16] and analyzing social media platforms like Twitter [8]. From there, Haines suggests that the next iteration for digital ethnographic work is the move to considering digital spaces as a site [9]. QE researchers have historically engaged in this way, using Twitter as a "place" to explore Covid discussions [8] and discussion forums as a site for game communities that support user identity change relevant to the gaming participatory culture [6].

The next step for digital ethnographic work as advocated by Haines is to move beyond the digital as the sole object of study and to consider it as a dimension of social life, embedded within the broader context of social and cultural practices. Viewing digital as a dimension surfaces new possibilities for understanding the interplay between online and offline worlds and is supported by the inclusion of interaction data. If the research around Twitter had included interaction logs, greater insight into the role of social media for isolated people may have been evident alongside the conversation content. Gameplay data may have allowed researchers to connect game-based achievements and events to the trajectory of community member identities. These added dimensions can support QE research by supporting interpretation of the phenomena at hand, and that interpretation begins with coding.

2.1 Relationship of Data to Codes

Data itself can be thought of as the recording of some observation. With digital data, the recording is the manifestation of an implicit conversation between the humans -- the designer and the user -- and the computer messenger. In order for these observations to be developed into a more comprehensive understanding of user activity, a stage of description must take place that allows for interpretation of the many events present within the logs. That a sensor tripped or a button clicked can only tell exactly *what* has happened [17]; codes encapsulate the meaning of a series of logged interaction events within the physical and digital context of the users.

The difficulty in digital data coding is that there is relatively little support to move beyond the *what*. In discourse coding, language plays a role in enabling interpretation by those who were outside of the initial conversation. While coding discourse transcripts gives primacy to certain elements (the literal words) over others (such as body language), language itself can generally be understood by the researchers as it was by the participants. This may not be the case when researchers seek to interpret raw logged events. Instead, there may be a process of transforming the data from the raw event logs, which are typically a form of shorthand developed by a programmer to represent user and system behavior, into something comprehensible by humans. There are several approaches being used to make meaning of this data through coding including participant co-examination of visualizations [18], video replay [19], and text replays [20]. In each case, researchers make sense of the logged events in aggregate by narrativizing the actions as a larger unit. While being able to watch the series of actions or read through them like a story is useful in exploring the data, this only provides the first step towards coding the data. The next step is determining how to operationalize a system for coding (a step taken in each of these approaches). It is with

this in mind that we consider the question: How should codes be operationalized for interaction log data from a quantitative ethnographic paradigm? To address this question, we examine the methods used by QE adjacent communities to operationalize their descriptive codes on data, both through the liberal use of evidence-centered definitions and increasing use of machine learning after an initial human interpretation step. We provide examples of these processes before discussing implications for broadening the QE toolkit to include machine learning in this way.

3 Codes from Evidence

Digital systems which have been designed using evidence-centered design principles [21] support the interpretation of user interactions into meaning. In such systems, interactions are intentionally tied to constructs, so each task a user partakes in can be directly connected to the component behaviors as individual pieces of evidence for or against a larger construct. In this way, the design of the system begins with the codes as the behaviors that designers wish to see from their users.

In Physics Playground, Karumbaiah et al. [3] use codes that describe simple machines to translate user behaviors -- drawing lines with varying lengths, curves, slopes, and connections -- into evidence for whether a player understood the challenge of a given puzzle. Similarly, user relationships to aggregate features (e.g. relative time spent on a puzzle compared with other players) were used as codes in discussing fairness of Shadowspect, a spatial reasoning game-based assessment [22]. The aggregate features connected theoretical understandings of persistence to the behavior observed across players within the digital, game-based context, thus demonstrating interaction codes' fairness to theory.

However, this approach also raises questions around the need to be fair to data and community [23]. Evidence-centered codes may not be enough to fully capture the variety of behaviors present in a system which manifest in the interaction data. In Physics Playground, questions arose around why players were quitting certain levels. The system was not intentionally designed to evoke quitting behavior as a means of gathering evidence on student learning; rather, instances of quitting were identified and observed during the examination of interaction logs [3]. Connecting the emergent quitting behavior to behaviors that the game was intentionally designed to elicit allowed researchers to better understand when students had shortcomings in their understanding of the utility of certain simple machines. However, quitting is a fairly straightforward behavior to code for on the basis of one event in the interaction log – players left a level without completing it successfully. Other behaviors may not be as straightforward to interpret from the events themselves. Thus, additional techniques may be necessary to map the emergent behaviors backwards to the interaction events and data features which characterize them.

4 Developing Emergent Codes from Interaction

Qualitative codes of interaction data can be used in two ways: as objects of analysis in and of themselves, or as the basis (training set) for a machine learning approach, much as qualitative codes of text are used to train tools such as nCoder [14]. The first task in either of these research paradigms is the same: generating human-coded data labels. In this step, researchers code video and text replays of interaction data to generate labeled datasets which can be analyzed directly or used as training data for machine learning classification tasks [24]. Recent work has attempted to partially automate this step as well, using artificial intelligence to suggest coding categories or conduct mixed-initiative coding [25], but this work is outside the scope of this paper.

In order to apply machine learning after the human coding step, it is necessary to distill features of the data that can support computer detection of the differences between code examples and nonexamples. This process typically involves human design of the features which are aggregates of elements of the captured interactions [26]. It is an iterative process to determine which features will be most effective at describing the codes as developed by the researcher [27], a process similar to the iterative code development described by Shaffer and Ruiz [23]. In this section, we elaborate on the processes used to both interpret and label interaction data as well as means by which researchers identify features that can be used to automate coding.

4.1 Creating a Dataset of Qualitative Codes

The coding process within and beyond QE relies on being able to create descriptions that are meaningful and interpretable beyond one person familiar with the data. This is the role of agreement metrics like kappa and Shaffer's rho for the QE community [28]. Working towards agreement through triangulation between human coders and machines allows for a minimization of uncertainty for the reliability of a description of a phenomena [23]. The trouble with interaction logs is that each event can be likened to a word in a sentence. There is potential for meaningful codes when the words come together, but it can be challenging to find complex meaning in a single word. Thus, when researchers seek to identify labels for the phenomena in interaction logs, they often utilize alternative representations (e.g. text, video, or visualizations) to assist human interpretation of the logged events.

Take, for example, text replays, used in many papers within EDM [20]. Text replays turn the events registered by the computer system (such as clicks) into textual descriptions of the event that can be read narratively. The process of translation does not require additional interpretation by the researcher as the descriptions provided can be taken directly from the data schema of the system designers. Using this method, the "observations" being recorded take the perspective of the data logging designers who may give primacy to certain types of user events and computer feedback. The strength of text replay is that it transforms the event stream into a story which can be segmented into different sized utterances. For example, in order to identify whether students struggle on a given task, it may be necessary to see the task from beginning to end; however, to identify productive use of guide text, it may only be necessary to

see the first few actions after a user receives help. Decisions about clip size, the EDM term for utterance, are therefore iteratively grounded in the behavior being identified and the data itself.

Video replays [19] are somewhat more removed from the interaction log itself, but they allow behavior to be situated within the context of the digital tool (as seen by the user). Video replays typically reconstruct logs into a movie rendition of the user interactions. Therefore, observation is less from the perspective of the designer and more from that of the user or traditional researcher lens. During coding, this allows the researcher to consider the state of the technical system, a particularly important element in games and other dynamic technologies, without the cognitive load of trying to track the state between system events because it is all viewable on the screen. Thus, coding can proceed more naturally with researchers considering the contextual game state as part of the determining factors for a given code.

Visualizations are one of the less transparent methods of coding interaction logs because there is less direct translation from the log itself to something observable by the researcher. Instead, actions are considered in aggregate as rates of change or quantified comparisons. For example, in their mapping of player activity in Plant Wars, researchers created metrics for the amount of fertilizer players were applying in game and plotted it based on the time the activity took place [29]. This mapping allowed them to easily answer questions around what and when behaviors were taking place, but it was harder to answer the question of why. Ultimately, a shift in when player activity was occurring encouraged researchers to dig deeper by interviewing players whose data was particularly representative of the anomaly of interest; these interviews elicited newfound understandings by identifying key context: some players were recent graduates whose sleep schedules had changed. Without the mappings to help surface questions, the researchers may never have been able to understand that element of gameplay behavior and misattributed the behavior to unrelated factors.

The ability to question the nature of the data beyond summary statistics is valuable because it allows research teams to ground the interpretation in cultural analysis, acknowledging the uniqueness of individual experience instead of normalizing assumptions about populations [18]. Visualizations can also be used to support participatory research which increases agency for participants in data collection and interpretation. Previously, QE researchers have utilized visualizations of a researcher-developed model to engage participants in participatory research [30], but participants may also be able to identify how trends connect with overarching codes from the initial stages of research as well. When HCI researchers were interested in understanding trends of device usage within the home, they used interaction logs alongside location data to help frame conversations with participants [17], an example of data-driven retrospective interviewing [31]. Engaging the participants in their own data allowed for the trends to be interrogated and framed by participants themselves (e.g. participants could define what family dinner looked like in their own data) [18].

Regardless of the method used to identify the codes and understand what is happening within the technical elements of the system, the plentiful nature of interaction logs necessitates automated coding if researchers desire to examine interactions beyond a few cases. Thus, we need a means for taking the insights and labels on our

dataset and creating interpretable features that computers can utilize for automating the detection of the codes within the larger corpus of data.

4.2 Determining Relevant Features

In discourse-centered QE research, many researchers have chosen to use the tool nCoder to assist in the automation of codes [14]. The tool supports human coders in calculating agreement (in service of validating codes) with a computer based on textual features of each utterance. The human coders provide the desired features for the computer in the form of regular expressions. For example, if a researcher is trying to capture people talking about symptoms, they may choose to include regular expressions like "headach*" and "feel" to try and capture the discussion. These expression-based features must be developed in relation to the corpus as a whole to avoid unintentionally indicating positive cases to the computer. The more data the researcher sees, the more refined they can make their features.

Similar approaches are used by EDM/LA researchers when working with interaction logs (see discussion in [11, 13, 32]). Once the labels have been established for positive and negative cases, they can be used to inform the feature generation process. However, features of interaction logs are not as straightforward as identifying a word that occurs in a sentence. Instead, aggregate features are calculated at the level of clip size to provide characteristics of the utterance that are interpretable by the computer. These features can then be used in a variety of algorithms to attempt to delineate between examples and nonexamples of the behavior.

One method to determine which features are relevant is to consider the human perspective of expertise. For example, Paquette et. al [33] considered the ways that experts thought through whether students were gaming the system in an intelligent tutoring system (ITS). The experts noted the importance of pause length for identifying the behavior within the interaction logs. When students in the ITS rapidly submitted answers over and over with little pause, experts deduced that the students were not taking time to try alternative strategies and were thus, likely gaming. These features may not be enough to support computer interpretation on their own. For example, Paquette and colleagues noted that the usefulness of the pause-based features is amplified by considering the edit distance (a metric that shows how different two submissions are) of each subsequent submission by the student. Gaming behavior could include rapid, formulaic answers (e.g. increasing subsequent answers by 10), thus the two features together assist computers in identifying the behavior.

Similar tactics could be utilized for code labels that originate from video replay or data visualization. Analogous to the identification of keywords as regex features, researchers need to consider what is leading them to code a given case positively. There are also general guidelines regarding what kinds of features can provide enough information for description of system interaction behavior [34]. Baker & Owen [34] describe the importance of including user behavior, system feedback, and user progression when generating feature sets. In the Plant Wars example [29], researchers designed features about how much fertilizer was being used, whether the game system indicated needing fertilizer, and player progress metrics.

Concerns may arise in thinking about features as aggregates (i.e. time between actions, number of clicks, etc.) that seem far removed from the phenomena of interest (i.e. whether a user is scanning for information or persisting in gameplay). This is why it is important to consider how each feature is relevant to the phenomena being observed. Grounding both label and feature generation in theory (see examples in [10, 35]) can be beneficial to addressing questions of validity. Feature selection techniques such as correlation and checking for collinearity may also allow researchers to filter out less meaningful features. Striking a balance between useful features that correlate with the phenomena and features based on a more traditional view of construct validity may be helpful in increasing transparency of the resultant model for stakeholders while maintaining the ability to computationally differentiate examples and nonexamples. To once again draw a comparison to textual data, it may be useful to incorporate a keyword "black box" in automating a code for "trust" in the context of AI discussions, but the keyword on its own may oversimplify the discussion. QE relies on being able to point back at the evidence for why an utterance is coded in a particular way and contextualize that decision [23]. Therefore, understanding the relationship between feature selection and the phenomena of interest, while providing explanations for their relevance, is crucial for ensuring fairness and contextually-grounded decisions in QE research with interaction data.

4.3 Making Models

Explainability is equally as important in model selection to detect code presence as it is in feature generation. Some ML models are more explainable or transparent than others in the way they present results, although recent work has attempted to increase the explainability of more inscrutable algorithms such as neural networks [36, 37].

There are potential challenges for QE researchers in trying to utilize AI models, even the more interpretable or explainable ones. For example, the degree to which explanations are interpretable to someone who is not an ML expert varies considerably [37]. The potential to automate the coding of identified behaviors offers a strong incentive to explore machine learning models for QE, enabling deeper exploration of these behaviors in connection with other behaviors and participant groups.

5 Why ML Codes are Useful for Understanding Interaction

In this section, we provide examples of ML codes and their utility for QE research to understand interaction data. We draw connections between the details that QE allows researchers to "get right" and the ways that the ML model would support those goals.

5.1 Situatedness Matters

Interactions emerge from "conversation" between participants and the digital system [4]. These sociotechnical systems are complicated, and the interactions are nuanced in a way that a click may not be just a click -- it may communicate a great deal more. To

understand the subtleties of interactions, researchers must be able to use their codes to "generalize within" the context to consistently describe interactions where the patterns of both user and computer system behavior remain consistent [38].

Take for example, the user behavior called Wheel-Spinning [39]. Students in digital learning environments may exhibit this behavior when they are stuck, thus continuing to try to solve a problem without making progress. Given just user actions, one may see a number of clicks and blank submissions, changes to strategy (such as asking the system for help), and even breaks between actions where users are considering new approaches. The system actions would likely be almost wholly consisting of negative feedback. Contextualizing the actions and negative feedback in light of a lack of forward progress allows for the existence of a Wheel-Spinning code, but so what?

QE and the focus on behaviors within context would allow exploration of this behavior within the classroom. Many educators would likely agree that wheel-spinning is an intervention-worthy behavior. Do the teachers successfully intervene? Do they even notice? Do other students notice? How does a student's Wheel-Spinning behavior manifest in discourse or collaboration? These are all questions that rely on a deeper understanding of the behavior in context. To answer these questions, we need to be able to connect the behaviors, not just clicks or feedback, that cross the digital boundaries to the situational context they exist within. It isn't just about being able to detect the behavior and tag it for interactions, but using QE, we may be able to uncover the nature of the system and its implications.

Furthermore, the use of detectors to code the interaction log may be a methodological necessity in this case. When participant behaviors cross over into digital spaces and spill into other physical spaces (either via transience in their own position or interaction with other participants who are not collocated), local observation by the ethnographer may not adequately capture the phenomena under observation [27]. In other words, if a researcher is watching a student in one classroom who is playing a collaborative game with students in another classroom, they may miss the potential downstream interactions happening with the other students as a result of the digital interactions because they are focused on the individual in front of them. The choices we make as researchers in what we choose to include and attend to as data impacts the way we return to the observations for analysis [40]. In an effort to not lose the metaphorical forest for the trees, tools (like SPACLE [41]) have been developed to assist researchers in extending their observations to position individual interaction logs alongside whole class data and observations in a replay system. By employing the use of logged interactions, we can consider a perspective that incorporates more elements of the sociotechnical system in situ.

5.2 Perspectives Matter

We must also consider whether the codes themselves consider the user in order to achieve the QE principle of fairness to the community [23]. Development of emic codes (those which are community generated) may come directly from the user during open-ended experience sampling [42], for instance. Users may participate in surveys

where questions attempt to uncover latent constructs such as affective states; however, these self-report measures can be unreliable when moving to interpersonal ratings and are weakly correlated to external behavioral measures [43]. Furthermore, self-report may be distorted by the user's ability to accurately describe their state, a potential problem for children, and in cases where the affect/behavior is not socially acceptable. An alternate approach, in-the-moment interviews driven by detectors, creates some potential to probe deeper than a survey can [44]. For example, being able to detect frustration or other negative affect through interaction logs can provide a way to ground conversations around how design choices impact user experience [44]. Similar to visualization exploration, interviews where users revisit their experience with the technical system allow for redefinition of codes. For instance, as users discuss their experience in interviews, their responses may lead the researcher to uncover nuances between feelings of frustration and challenge, prompting a need to revise the code model for a more accurate representation.

Why rely on ML models to begin these conversations? Tacit knowledge is often left out of conversations and interviews with participants; researchers don't know to ask the questions, and the actions are so intuitive to users that they may not be readily articulated to an observer [27]. Assigning labels to moments of interest from the perspective of the researcher allows for more in-depth questioning about the perspective of the participant.

6 Unknown Compromises in Using Machine Learning

Although there are several potential benefits to the use of machine learning coding of interaction data, there are several compromises that are inherent in incorporating this practice into the QE workflow. Principally, there is a need for discussions centered on the ideas of triangulation and transparency. We begin the discussion here, raising questions and providing guidance for QE researchers looking to join this trajectory.

6.1 Triangulation

Interaction logs are not the only form of "big data," and automated coding outside of detectors has been heavily leveraged by the QE community through tools like nCoder [14]. These tools require not only agreement between the human coders on training data, but also agreement between the automated classifier and both human coders in order to minimize uncertainty [23]. This is also a standard practice when developing detectors in digital environments. In almost all cases, a machine-learned model's performance is evaluated on whether its inferences agree with human coders for data not used to develop the models. In exemplary cases, when generalizability beyond the initial data and context is imperative, after a machine-learned model has been fully completed, entirely new data is obtained, coded by humans, and tested for agreement.

When models underperform, researchers often return to the data in attempts to improve them by both inspecting areas where models mislabeled the data and creating new features to help capture previously missing elements [45]. Each subsequent re-

turn to the data allows researchers to refine their understanding of the relationship between features of the data and the code. The developers of Aeonium, an ML tool that supports qualitative coding, recommend embracing this practice and especially the ambiguity that arises from disagreement as a part of the process, bringing forth subjectivity and data inconsistencies for interrogation leading to better codes [46].

Additionally, discussions for acceptable kappa thresholds will be required within the QE community. Traditionally, kappa values above .61 have been seen as acceptable in other communities [47]. In QE community tools, kappa levels of .9 are the expectation [14]. In EDM and LA, much lower levels of kappa -- as low as .2 -- are sometimes seen as acceptable, depending on the way a model will be used. For example, even slightly better than chance performance may be valuable if the potential benefit of an intervention is high and the cost of an incorrectly-delivered intervention is low. These differences require that researchers are explicit about the strengths and weaknesses of their ML code models and the models that the codes are introduced into, such as network analyses. For machine-learned models, whether developed using nCoder or on interaction data, careful consideration is needed of whether the model's accuracy is sufficient for its intended use, and what the risks to interpretation and fairness are of using an imperfect model.

6.2 Transparency

A component of the detailed explanation expected of QE researchers is the transparency with which they can tell the story of their data. Transparent models are essential for closing the interpretive loop [28] by connecting the final model to the codes and back to the data features that labeled the interaction with that code. Different ML models have different levels of explicability and transparency. Recent work has attempted to increase the explainability of even inscrutable models such as neural networks [37], but these attempts remain imperfect, and explainability methods often disagree with each other [48]. This reifies the trade-off within machine learning between accuracy and explainability -- more accurate models are often harder to understand than simpler, less accurate models. This challenge is not unique to interaction data; contemporary NLP approaches involving large language models can often produce much better performance (particularly for unseen data) but are much harder to explain or interpret. Algorithm selection may come down to weighing the cost of poorer accuracy against the cost of less transparency and interpretability.

7 Conclusions

Qualitative interpretations of a phenomena are deeply personal to the position of the researcher, the participants, and the context within which the observations take place [49]. Digital tools are now pervasive in the spaces we work in, and many phenomena of interest to QE researchers also involve digital data. These tools offer potential insights that may help to reframe participant actions and behaviors that cannot be fully understood from place-based observation or even participant interviews. Leveraging

data as a talking point with participants may allow researchers to identify novel behaviors and better understand how those behaviors are situated within the context of study (e.g. life changes for young gamers [29]). Machine-learned models developed using interaction data may provide insights into fine-grained aspects of user activity that are emergent and hard to otherwise study. Interaction logs provide valuable information to the QE researcher; the challenge for the community will be to continue finding and refining methods that allow for transparent description of this data and to continue pursuing development and alignment of QE paradigms for research involving this data, while maintaining the principles and research values of the QE community more broadly.

8 References

1. Pentina, I., Hancock, T., Xie, T.: Exploring relationship development with social chatbots: A mixed-method study of replika. Computers in Human Behavior. 140, 107600 (2023). https://doi.org/10.1016/j.chb.2022.107600.
2. Shah, M., Siebert-Evenstone, A., Moots, H., Eagan, B.: Quality and Safety Education for Nursing (QSEN) in Virtual Reality Simulations: A Quantitative Ethnographic Examination. In: Wasson, B. and Zörgő, S. (eds.) Advances in Quantitative Ethnography: Third International Conference, ICQE 2021, Virtual Event, November 6-11, Proceedings. pp. 237–252. Springer (2022).
3. Karumbaiah, S., Baker, R., Barany, A., Shute, V.: Using Epistemic Networks with Automated Codes to Understand Why Players Quit Levels in a Learning Game. In: Eagan, B., Misfeldt, M., and Siebert-Evenstone, A. (eds.) Advances in Quantitative Ethnography: First International Conference, ICQE 2019, Madison, WI, USA, October 20–22, 2019, Proceedings. pp. 106–116. Springer (2019).
4. Scianna, J., Gagnon, D., Knowles, B.: Counting the Game: Visualizing Changes in Play by Incorporating Game Events. In: Ruis, A.R. and Lee, S.B. (eds.) Advances in Quantitative Ethnography. pp. 218–231. Springer International Publishing, Cham (2021).
5. Scianna, J., Kaliisa, R., Boisvenue, J., Zörgő, S.: Approaching Structured Debate with Quantitative Ethnography in Mind. In: Wasson, B. and Zörgő, S. (eds.) Advances in Quantitative Ethnography: Third International Conference, ICQE 2021, Virtual Event, November 6-11, Proceedings. pp. 33–58. Springer (2022).
6. Barany, A., Foster, A.: Examining Identity Exploration in a Video Game Participatory Culture. In: Eagan, B., Misfeldt, M., and Siebert-Evenstone, A. (eds.) Advances in Quantitative Ethnography: First International Conference, ICQE 2019, Madison, WI, USA, October 20–22, 2019, Proceedings. pp. 3–13. International Society for Quantitative Ethnography (2019).
7. Arastoopour Irgens, G.: Using Knowledgeable Agents of the Digital and Data Feminism to Uncover Social Identities in the #blackgirlmagic Twitter Community. Learning, Media and Technology. 47, 79–94 (2022).
8. Hobbs, W., Kaliisa, R., Misiejuk, K., Peffer, M., Sanchez, D., Scianna, J., Shah, M., Talafian, H., Vachuska, K., Wang, Y., Wang, S., Woodard, M.A., Zorgo, S.: Challenges and Solutions to Examining Twitter Data: Reflections from QE-

COVID19 Data Challenge. In: Second International Conference on Quantitative Ethnography: Conference Proceedings Supplement. p. 84. , Online (2021).
9. Haines, J.K.: Towards Multi-Dimensional Ethnography. Ethnographic Praxis in Industry Conference Proceedings. 2017, 127–141 (2017). https://doi.org/10.1111/1559-8918.2017.01143.
10. Gašević, D., Dawson, S., Siemens, G.: Let's not forget: Learning analytics are about learning. TechTrends. 59, 64–71 (2015).
11. Reimann, P., Frerejean, J., Thompson, K.: Using process mining to identify models of group decision making in chat data. In: O'Malley, C., Suthers, D., Reimann, P., and Dimitracopoulou, A. (eds.) Computer Supported Collaborative Learning Practices: CSCL2009 Conference Proceedings. pp. 98–107. International Society of the Learning Sciences, Rhodes, Greece (2009). https://doi.org/10.22318/cscl2009.1.98.
12. Baker, R.Sj., Ocumpaugh, J.: 16 Interaction-Based Affect Detection in Educational Software. The Oxford handbook of affective computing. 233 (2014).
13. Bruckman, A.: Analysis of log file data to understand behavior and learning in an online community. The International handbook of virtual learning environments. 1449–1465 (2006).
14. Marquart, C., Swiecki, Z., Eagan, B., Shaffer, D.W.: ncodeR: Techniques for automated classifiers [R package], https://cran.r-project.org/web/packages/ncodeR/index.html, (2019).
15. Shaffer, D.W., Hod, Y.: The Role of Ethnography in Quantitative Ethnography. In: Fourth International Conference on Quantitative Ethnography: Conference Proceedings Supplement (2022).
16. Kozinets, R.V., Gretzel, U.: Netnography. In: Encyclopedia of Tourism Management and Marketing. pp. 316–319. Edward Elgar Publishing (2022).
17. Aipperspach, R., Rattenbury, T., Woodruff, A., Anderson, K., Canny, J., Aoki, P.: Ethno-Mining: Integrating Numbers and Words from the Ground Up. 13 (2006).
18. Anderson, K., Nafus, D., Rattenbury, T., Aipperspach, R.: Numbers Have Qualities Too: Experiences with Ethno-Mining. Ethnographic Praxis in Industry Conference Proceedings. 2009, 123–140 (2009). https://doi.org/10.1111/j.1559-8918.2009.tb00133.x.
19. Harpstead, E., MacLellan, C.J., Aleven, V., Myers, B.A.: Replay analysis in open-ended educational games. Serious games analytics: Methodologies for performance measurement, assessment, and improvement. 381–399 (2015).
20. Baker, R., de Carvalho, A.: Labeling student behavior faster and more precisely with text replays. In: Educational Data Mining 2008 (2008).
21. Mislevy, R.J., Almond, R.G., Lukas, J.F.: A Brief Introduction to Evidence-Centered Design. ETS Research Report Series. 2003, i–29 (2003). https://doi.org/10.1002/j.2333-8504.2003.tb01908.x.
22. Kim, Y., Choi, J.: Expanding Fairness in Game-Based Assessment with Quantitative Ethnography. Presented at the International Conference on Quantitative Ethnography 2022 , Copenhagen, Denmark (2022).
23. Shaffer, D.W., Ruis, A.R.: How We Code. In: Ruis, A.R. and Lee, S. (eds.) Advances in Quantitative Ethnography: Second International Conference, ICQE

2020, Malibu, CA, USA, February 1-3, 2021, Proceedings. pp. 62–77. Springer (2021).
24. Zhang, J., Andres, J.M., Alexandra, L., Hutt, S., Baker, R.S., Ocumpaugh, J., Mills, C., Brooks, J., Sethuraman, S., Young, T.: Detecting SMART Model Cognitive Operations in Mathematical Problem-Solving Process. International Educational Data Mining Society. (2022).
25. Choi, J., Ruis, A.R., Cai, Z., Eagan, B., Shaffer, D.W.: Does Active Learning Reduce Human Coding?: A Systematic Comparison of Neural Network with nCoder. In: Damşa, C. and Barany, A. (eds.) Advances in Quantitative Ethnography: Fourth International Conference, ICQE 2022, Copenhagen, Denmark, October 15–19, 2022, Proceedings. pp. 30–42. Springer (2023).
26. Jiang, Y., Bosch, N., Baker, R.S., Paquette, L., Ocumpaugh, J., Andres, J.M.A.L., Moore, A.L., Biswas, G.: Expert feature-engineering vs. deep neural networks: which is better for sensor-free affect detection? In: Artificial Intelligence in Education: 19th International Conference, AIED 2018, London, UK, June 27–30, 2018, Proceedings, Part I 19. pp. 198–211. Springer (2018).
27. Hsu, W.F.: Digital ethnography toward augmented empiricism: A new methodological framework. Journal of Digital Humanities. 3, 3–1 (2014).
28. Shaffer, D.W.: Quantitative Ethnography. Cathcart Press (2017).
29. Shadoan, R., Dudek, A.: Plant Wars Player Patterns: Visualization as Scaffolding for Ethnographic Insight, http://ethnographymatters.net/blog/2013/04/11/visualizing-plant-wars-player-patterns-to-aid-ethnography/, last accessed 2023/04/23.
30. Vega, H., Arastoopour Irgens, G.: Constructing Interpretations with Participants Through Epistemic Network Analysis: Towards Participatory Approaches in Quantitative Ethnography. In: Wasson, B. and Zörgő, S. (eds.) Advances in Quantitative Ethnography: Third International Conference, ICQE 2021, Virtual Event, November 6-11, Proceedings. pp. 3–16. Springer (2022).
31. El-Nasr, M.S., Durga, S., Shiyko, M., Sceppa, C.: Data-driven retrospective interviewing (DDRI): a proposed methodology for formative evaluation of pervasive games. Entertainment Computing. 11, 1–19 (2015).
32. Baker, R.Sj., Corbett, A.T., Wagner, A.Z.: Human classification of low-fidelity replays of student actions. In: Proceedings of the educational data mining workshop at the 8th international conference on intelligent tutoring systems. pp. 29–36 (2006).
33. Paquette, L., de Carvalho, A.M., Baker, R.S.: Towards Understanding Expert Coding of Student Disengagement in Online Learning. In: CogSci (2014).
34. Owen, V.E., Baker, R.S.: Fueling Prediction of Player Decisions: Foundations of Feature Engineering for Optimized Behavior Modeling in Serious Games. Tech Know Learn. 25, 225–250 (2020). https://doi.org/10.1007/s10758-018-9393-9.
35. Beigi, G., Tang, J., Liu, H.: Social Science Guided Feature Engineering: A Novel Approach to Signed Link Analysis. ACM Trans. Intell. Syst. Technol. 11, 1–27 (2020). https://doi.org/10.1145/3364222.
36. Xu, F., Uszkoreit, H., Du, Y., Fan, W., Zhao, D., Zhu, J.: Explainable AI: A brief survey on history, research areas, approaches and challenges. In: Natural Language Processing and Chinese Computing: 8th CCF International Conference,

NLPCC 2019, Dunhuang, China, October 9–14, 2019, Proceedings, Part II 8. pp. 563–574. Springer (2019).
37. Angelov, P., Soares, E.: Towards explainable deep neural networks (xDNN). Neural Networks. 130, 185–194 (2020).
38. Munk, A.K., Olesen, A.G., Jacomy, M.: The Thick Machine: Anthropological AI between explanation and explication. Big Data & Society. 9, 1–14 (2022). https://doi.org/10.1177/20539517211069891.
39. Beck, J.E., Gong, Y.: Wheel-spinning: Students who fail to master a skill. In: Artificial Intelligence in Education: 16th International Conference, AIED 2013, Memphis, TN, USA, July 9-13, 2013. Proceedings 16. pp. 431–440. Springer (2013).
40. Hall, R.: Videorecording as Theory. In: Handbook of Research Design in Mathematics and Science Education. pp. 647–64. Lawrence Erlbaum, Mahwah, NJ (2000).
41. Holstein, K., McLaren, B.M., Aleven, V.: SPACLE: investigating learning across virtual and physical spaces using spatial replays. In: Proceedings of the Seventh International Learning Analytics & Knowledge Conference. pp. 358–367 (2017).
42. Zirkel, S., Garcia, J.A., Murphy, M.C.: Experience-Sampling Research Methods and Their Potential for Education Research. Educational Researcher. 44, 7–16 (2015). https://doi.org/10.3102/0013189X14566879.
43. Dang, J., King, K.M., Inzlicht, M.: Why Are Self-Report and Behavioral Measures Weakly Correlated? Trends in Cognitive Sciences. 24, 267–269 (2020). https://doi.org/10.1016/j.tics.2020.01.007.
44. Baker, R.S., Nasiar, N., Ocumpaugh, J.L., Hutt, S., Andres, J.M.A.L., Slater, S., Schofield, M., Moore, A., Paquette, L., Munshi, A., Biswas, G.: Affect-Targeted Interviews for Understanding Student Frustration. In: Roll, I., McNamara, D., Sosnovsky, S., Luckin, R., and Dimitrova, V. (eds.) Artificial Intelligence in Education. pp. 52–63. Springer International Publishing, Cham (2021).
45. Slater, S., Baker, R.S., Wang, Y.: Iterative Feature Engineering through Text Replays of Model Errors. International Educational Data Mining Society. (2020).
46. Chen, N.-C., Drouhard, M., Kocielnik, R., Suh, J., Aragon, C.R.: Using Machine Learning to Support Qualitative Coding in Social Science: Shifting the Focus to Ambiguity. ACM Trans. Interact. Intell. Syst. 8, 1–20 (2018). https://doi.org/10.1145/3185515.
47. McHugh, M.L.: Interrater reliability: the kappa statistic. Biochemia medica. 22, 276–282 (2012).
48. Swamy, V., Radmehr, B., Krco, N., Marras, M., Käser, T.: Evaluating the explainers: black-box explainable machine learning for student success prediction in MOOCs. arXiv preprint arXiv:2207.00551. (2022).
49. O'Connor, C., Joffe, H.: Intercoder Reliability in Qualitative Research: Debates and Practical Guidelines. International Journal of Qualitative Methods. 19, 1609406919899220 (2020). https://doi.org/10.1177/1609406919899220.