Detector-Driven Classroom Interviewing: Focusing Qualitative Researcher Time by Selecting Cases *in Situ*.

Ryan S. Baker[a], Stephen Hutt[a], Nigel Bosch[b], Jaclyn Ocumpaugh[a], Gautam Biswas[c], Luc Paquette[b], J. M. Alexandra Andres[a*], Nidhi Nasiar[a], and Anabil Munshi[c]

[a]Graduate School of Education, University of Pennsylvania, Philadelphia, USA; [b]School of Information Sciences, University of Illinois Urbana-Champaign, Champaign, USA; [c]School of Engineering, Vanderbilt University, Nashville, USA

Abstract

In this paper, we propose a new method for selecting cases for *in situ*, immediate interview research: Detector-Driven Classroom Interviewing (DDCI). Published work in educational data mining and learning analytics has yielded highly scalable measures that can detect key aspects of student interaction with computer-based learning in close to real-time. These measures detect a variety of constructs and make it possible to increase the precision and time-efficiency of this form of research. We review four examples that show how the method can be used to study why students become frustrated and how they respond, how anxiety influences how students respond to frustration, how metacognition interacts with affect, and how to improve the design of an adaptive learning system. Lastly, we compare DDCI to other mixed-methods approaches and outline opportunities for detector-driven classroom interviewing in research and practice, including research opportunities, design improvement opportunities, and pedagogical opportunities for teachers.

Keywords: detector-driven classroom interviewing, interview research, mixed methods, computer-based learning

Detector-Driven Classroom Interviewing: Focusing Qualitative Researcher Time by Selecting Cases *in Situ*

Qualitative research, in its various forms, is of great importance to modern educational research (Ravitch & Carl, 2019). Qualitative research enables the in-depth investigation of phenomena of interest, and yields key insights not just on what happens, but when used effectively, why phenomena are occurring.

Among the considerable variety of qualitative methods, one of the most important categories is interview methods. Interviews give a researcher an opportunity to capture information about a phenomenon in terms of the participant's own way of thinking and referring to the phenomenon (Kvale & Brinkmann, 2009). In addition, interviews enable a researcher to probe into participant thinking in greater depth; when a participant refers to something, the interviewer can ask for greater detail and explanation. Whether the interviewer situates themselves with regards to the participant in an ethnographic fashion (Spradley, 2016) or more as an outsider, the ability to ask follow-up questions gives researchers the ability to check their own understanding and to obtain key detail that could be missed in methods where the protocol is set in advance and cannot be adjusted.

Current interview techniques can be categorized in many ways; one key division is whether the interview occurs during the activity or retrospectively. Retrospective interviews allow researchers to investigate participants' holistic experience with the activity and to investigate their memories of a few specific cases, but usually just those that are easily remembered (van Someren et al., 1994). These interviews are unlikely to capture transient cognitive and emotional experiences, the memory of

which is likely to change as the context and goals of the interviewee shift (Barrett, 2017).

Memories of earlier activities are also likely to be influenced by social interactions that took place between those activities and the interview, as the interviewee processes their experiences through language and additional context (see Lindquist et al.'s (2006) discussion of semantic memory). For example, consider the experience of being startled by a loud bang behind you in a classroom. In the moment, you might process that experience as quite disruptive and slightly terrifying, regardless of its origin. Yet, if you were interviewed two days later about the event, you would probably describe it quite differently depending on whether the bang was created by a birthday balloon popping or something more sinister. Past experiences are re-assessed over time, and later cognition about those experiences may not represent the initial way an individual looked at their experience. For example, a retrospective interview with two students who were surprised and confused by a new piece of knowledge might describe it differently if one student resolved that confusion while the other never fully assimilated the new information. Jerolmack & Kahn (2014) therefore argue that interviews are less useful when they are divorced from the context that is being investigated, since researchers have less access to direct evidence about the constructs they are studying when it is filtered through the interviewee's memory. In some cases, then, it can be advantageous to capture what students are thinking during the learning process—before they have time to process and recategorize their experiences. In some research communities it has therefore become standard to conduct interviews *in situ* (Hunting, 1997; Erbas & Okur 2012).

However, interviews occurring *in situ*, in the midst of the activity, have historically had several challenges. The first challenge is that the activity of interest may be

interrupted or hampered. This problem is not unique to interviewing; even think-aloud protocols can disrupt some types of activity and cognition (Schooler et al., 1993). A second challenge, more serious in some contexts than others, is that the researcher's presence may itself change the activity or influence the students' perceptions of it (Ward, 1981). Researchers in sociolinguistics, for example, have discussed this "observer's paradox" (Labov, 1972), and how to address it (also see Briggs, 1986; Seidman, 2006).

A third challenge can be described as the "needle in a haystack" problem. If the researcher wants to interview a participant about a phenomenon of particular interest, and this phenomenon is intermittent or occasional, the researcher may spend considerable time simply waiting for the phenomenon to happen. Worse yet, in a busy classroom with twenty or more students, the phenomenon may occur on one side of the room when the researcher is on the other. The researcher may completely miss the occurrence of the phenomenon they want to study.

This is particularly important when studying many key phenomena of interest to the modern learning sciences, such as self-regulated learning, metacognition, and affect, where critical events occur rarely. For example, transitions in learner affect may only occur once a minute (e.g., D'Mello & Graesser, 2011; Botelho et al., 2018), and many affective states of interest only occur 5-10% of the time (D'Mello, 2013), meaning an affective transition of interest may only occur 5-10 times across all students in a 45-minute class session. However, despite their rarity, some affective transitions have disproportionately strong associations with learning outcomes (Andres et al., 2019). This challenge either drives interview research to studying more frequently occurring events or requires months of being embedded in classrooms in order to study the most interesting phenomena (cf. Schofield, 1995).

A fourth challenge for *in situ* interview methods is in selecting cases and individuals for interviews. Current practice strives to avoid simply using convenience samples (Saldana, 2011; Emmel, 2013) and attempts to make sampling purposeful in nature (Ravitch & Carl, 2019), selecting cases and individuals who help shed light on the phenomena of interest. Patton (2002) and Emmel (2013) review strategies for purposeful sampling in detail.

For example, maximum variation sampling strategies try to find cases that are as different from each other as possible, intensity sampling tries to find cases that most intensely demonstrate the phenomenon being investigated, and typical sampling tries to find cases that demonstrate a typical case of the phenomenon (Patton, 2002; Emmel, 2013). Quantitative information can enter into the process of selecting cases within criterion sampling, where specific criteria are used to select cases for study. Researchers in a variety of fields have also recently begun to use data mining approaches to select cases for qualitative investigation (Luo, 2015; Hoeber et al., 2016). However, these methods have not yet been applied to selecting cases for *in situ* research, where it is important not just to select a case for in-depth study but to select the right time to conduct an interview.

Detector-Driven Classroom Interviewing (DDCI)

In this paper, we propose a new method for selecting cases for *in situ*, immediate interview research: Detector-Driven Classroom Interviewing (DDCI). DDCI attempts to address the third and fourth challenges to classroom interviewing discussed above. This method does not address all challenges to classroom interviewing, but instead focuses on the challenges of selecting which individual to interview and when. In this fashion, DDCI facilitates learning about uncommon but impactful events during real-world learning.

DDCI builds upon work over the last fifteen years in educational data mining and learning analytics (Baker & Siemens, 2014), which has yielded new, highly scalable, immediate measures of student interaction within the context of computer-based learning. These measures can detect, almost in real-time (with a delay under 30 seconds), a considerable variety of constructs, including a student's self-regulated learning strategies (Biswas et al., 2017), behaviors related to metacognition (Azevedo et al., 2010; Roll et al., 2011), affective state (Hutt et al., 2019), and disengaged behaviors (Baker & Rossi, 2013). These measures have been developed for a wide variety of computer-based learning environments, many used regularly in classrooms, including intelligent tutoring systems, games, simulations, exploratory and open-ended learning environments, and first-person virtual reality environments. Modern examples of these measures, developed using a combination of ground truth measures (such as classroom observations; Baker et al., 2020) and machine learning/data mining, often have high degrees of agreement to expert human judgment, and have been used both in research (see, for instance, Azevedo & Gasevic, 2019; Botelho et al., 2018; Hershkovitz et al., 2014) and to drive interventions which improve learning outcomes (see, for instance, D'Mello et al., 2010; DeFalco et al., 2018).

These measures offer an opportunity for classroom interview research as well—a solution to the needle in a haystack problem and to selecting individuals for interviews. Take, for instance, a researcher interested in a self-regulated learning strategy or affect transition that does not occur frequently. In DDCI, automated detectors can continually analyze the data being generated by students using learning software, wait for a set of key events of interest to occur, and then immediately notify the classroom interviewer when one of those events occurs. Rather than simply watching for an event of interest to occur, or perhaps

systematically (or haphazardly) interviewing students and hoping the event of interest will emerge, the researcher can focus their time on the event they want to study. This is especially important in cases where it is difficult to see an event of interest just by watching the student from a distance—for example, a student adopting a new self-regulated learning strategy in their use of a learning system.

A researcher using the DDCI approach can specify, before a class session, what events they are interested in researching, and then receive alerts when those events occur for specific students. They can then interview those students a minute after the event occurred—or in some cases, even more rapidly. While this may lower the likelihood of serendipitous discovery (Knapp, 1997), it shortens the time needed to obtain several interviews about a phenomenon of interest from weeks to a single class period.

Illustrating Detector-Driven Classroom Interviewing

We illustrate detector-driven classroom interviewing in the context of a program of research conducted within the computer-based learning environment, Betty's Brain. Betty's Brain (discussed in detail below) affords complex self-regulated learning behaviors; our goal was to study how self-regulated learning and affect interact with each other. We present below a detailed description of how we set up our research approach and infrastructure. We follow this with four vignettes from our research, previously published in more extended form in conference papers (Andres et al., 2022; Baker et al., 2021; Bosch et al., 2021; Ocumpaugh et al., 2021). These vignettes demonstrate the use of these methods to investigate a range of research topics:

- How student perceptions drive the experience of frustration during learning and in turn their choices during learning.

- The interplay between students' science anxiety, their affect, and their subsequent actions
- Understanding how students' metacognitive expressions correlate to their metacognition related behavior and their affective experiences
- Understanding how the design of the learning system's messages impacted their perceptions

Materials and Methods

Context: Betty's Brain

Betty's Brain is an open-ended, computer-based learning system that uses a learning-by-teaching paradigm to teach complex scientific processes (Leelawong & Biswas, 2008). Betty's Brain is primarily targeted towards middle school students, and students usually spend approximately one week per topic working individually with Betty's Brain as part of their regular classroom activities. Students typically complete two topics using the system within a school year. Teachers are also present during use and answer student questions and give suggestions.

Betty's Brain requires students to teach Betty, i.e., a virtual agent, about scientific phenomena (e.g., climate change, ecosystems, thermoregulation). In order to teach Betty, the student reads material about the phenomena and uses that information to construct a concept map that demonstrates the causal relationships between the various concepts related to that phenomena (see Figure 1).

Figure 1. Screenshot of viewing quiz results and checking the chain of links Betty used to answer a quiz question.

The learning process required by Betty's Brain necessitates high levels of self-regulation. As students construct their map, they must navigate through multiple hypermedia pages where they can read about a variety of concepts and their relationships. They can choose how often to test Betty's knowledge, and they may elect to interact with a virtual mentor agent (Mr. Davis) if they are having trouble teaching Betty. At times, the Mr. Davis may recognize student difficulties and intervene by suggesting strategies that the students can use to improve their performance in teaching Betty. The design of the system and the students' interactions with the agents requires and helps them to develop self-regulated learning skills to be successful in accomplishing their tasks in the Betty's Brain.

In other words, the pedagogical agents (Betty and Mr. Davis) provide a social framework for the gradual internalization of effective learning behaviors. An emphasis on self-regulatory feedback that has been demonstrated to lead to more effective self-regulated learning behaviors among students who use Betty's Brain, as well as better learning outcomes (Leelawong & Biswas, 2008).

Automated Detectors of Student Affect and Behavior

As students interact with Betty's Brain, automated detectors embedded in the learning system identify key moments and transitions in students' learning processes. These detectors can identify shifts between educationally relevant affective states such as boredom and confusion (Jiang et al., 2015) and behavioral sequences, such as deleting large sections of correct entries in the concept map after being notified that a part of the map is incorrect (Munshi et al., 2018). The affect and behavioral sequence detectors are continually running within Betty's Brain, analyzing the students' activities and their performance (i.e., their progress in building the correct map). After each student action, their behavior is distilled into

a set of features (including the type of action, such as adding a link in the causal map or clicking on the virtual mentor agent; also including higher-level features such as the time between actions and how many actions of the same type occur in a row). The affect detection looks for specific combinations of these features that are associated with specific affective states (Jiang et al., 2015). These combinations of actions are represented using logistic regression or step regression models. The combinations of actions were determined by first collecting hundreds of classroom observations of student affect, then using machine learning methods to discover which combinations of features corresponded to the classroom observations, and finally testing the detectors for validity on unseen data (Jiang et al., 2015).

A range of algorithms were tested, and logistic regression/step regression performed best on unseen data. The detectors outputted a probability for each of five affective states (boredom, frustration, confusion, engaged concentration, and delight) at each time point (using a 20-second grain-size). Affective transitions (from frustration to boredom, for instance), were identified when the detectors shifted from inferring which affective state was highest probability, from one state to another state.

The behavioral sequences were originally identified using sequential pattern mining (Munshi et al., 2018). Frequent patterns emerging from a sequence of logged learner activities were interpreted using a task model (Biswas et al., 2017) to identify cognitive-metacognitive behaviors related to learning processes such as information acquisition, solution construction and solution assessment. The behavior detection process represented the learner's current task context, their recent activities, and the effectiveness and coherence of these activities. Ineffective behavioral transitions would suggest moments of learning difficulties (Munshi et

al., 2023) where the student may be interviewed to gain a deeper understanding of their internal metacognitive (monitoring and self-reflection) processes.

APP: A Tool for Informing Researchers in Real Time

The detectors were used to prompt student interviews. Interviewers were signaled through a field research app, Quick Red Fox (QRF) (citation redacted), shown in Figure 2, which receives notifications from the detectors in Betty's Brain and allows users to record metadata related to each event (in this case, timestamps and which student was being interviewed). Technical and design details for the QRF app are provided in (Hutt et al., 2022). The specific affective sequences or behaviors that trigger an interview recommendation are selected by the researcher and can differ between studies. Examples of interview triggers are given in the vignettes below.

A prioritization algorithm selects which student should be interviewed in instances where multiple students display interesting patterns at roughly the same time. Prioritization is given to students who have not yet been interviewed (or who have not been recently interviewed), and in terms of a researcher-selected priority order for phenomena. If interviewers are not comfortable interrupting a student for any reason, they can skip the prompt within the app, and receive another recommendation from QRF.

Figure 2. The Quick Red Fox app for classroom interviewing.

How Interviews Were Conducted

As students worked on Betty's Brain, two interviewers in the classroom received information about specific affective or behavioral transitions in the QRF app

installed on their handheld mobile devices. Interviewers used this information to approach a student and speak with them about their learning process and strategies. Following a long line of research about how perceptions of interviewers affect people's conversational interactions (Goffman, 1959; Barriball & While, 1994), we noted that interviewers were most likely to be perceived by these students as being similar to either teachers or software developers, both of which entail power differences that could be counterproductive to obtaining truthful accounts of students' learning experiences (Eder & Fingerson, 2002; Hunting, 1997). Based on Wengraff's (2001) advice that the interview questions are only one component of a successful interview, interviewers attempted to take a helpful but non-authoritative role when speaking with students, situating themselves as friendly and sympathetic adults rather than arbiters of the right way to do things.

The interviews were intentionally brief (almost always under 5 minutes, and sometimes under a minute), in order to focus on the current situation and a small number of questions about it, and then let the student return to their work. Although the interviews were brief, the availability of real-time data on the student's recent behavior and the visibility of the student's current work on the screen made it possible to investigate a specific student decision or event in depth even in this short time. The interviews were conducted without a set script; however, they often started by asking students what strategies they are using (if any) for progressing in their work on the system. This approach was chosen for several reasons. First, it was a conversation starter that was likely to feel familiar to the students (thus setting them at ease at the start of the interview). Second, it was a conversation opener that would be appropriate across a wide range of interview triggers—both affective and metacognitive. Third, it allowed successful students the opportunity to present their own expertise about the system while being non-threatening to any students who

might still be struggling. Finally, it gave students an opportunity for reflection, potentially providing additional useful information about the students' reasoning behind their strategies.

As new patterns and information emerged in these open-ended interviews, the interviewer asked follow-up questions about the student's statements. For example, students who appeared to be struggling could be prompted to explain what they had tried and why they thought it might not be working. Meanwhile, students who were experiencing success were prompted to explain what they intended to do next.

Overall, students are encouraged to provide feedback about their experience with the software and talk about their choices as they navigate the platform. During these conversations, interviewers looked for general trends that could be explored. An example of this is given in our fourth interview vignette below, where we followed up initial comments that a software agent was rude in order to understand what aspect of the agent design led to this perception. In general, students were encouraged to offer their own opinions and to be forthcoming about any thoughts or emotions they experienced while interacting with the system.

In the following subsections, we discuss four vignettes that illustrate the use of DDCI for different research goals, in combination with a range of other methods. The goal of these vignettes is to illustrate the scope of research that DDCI can support within a single overall project.

Situating our Interview Technique within The Existing Literature

There is a large literature on different interview strategies. The approach used in our interviews has some commonalities with *semi-structured* interview methods (Kallio et al., 2016), which allow flexibility to change interview direction and give some degree of initiative to the interviewee while keeping to a pre-selected set of

themes. Although our interviews use data to select which themes to focus on, the interviews we have conducted thus far using QRF app can perhaps better be seen as *open-ended* rather than *semi-structured* because we did not typically immediately ask questions specific to the trigger. We did this to avoid biasing a student's responses by asking questions in a way that presupposed our perspective on their experience. In particular, we wanted to avoid asserting or assuming a specific lens on the student's recent experience; instead, we started by trying to understand how students understood their recent activity and experience. This approach to conducting each interview aligns with many viewpoints on Grounded Theory (see review in Aldiabat & Le Navenec, 2018), which recommend starting interviews in unstructured ways with the goal of focusing on the categories that emerge as salient (Strauss & Corbin, 1990). During the interviews, we also avoided making very specific labeling statements based on the triggers, in case students saw their experience different than the detectors, and also in case the detectors were inaccurate in this specific case.

While individual interviews were largely unstructured, our overall technique mirrors some aspects of semi-structured interviews. Since we might interview the same student several times, even multiple times within the same day, and across these sessions, the interviewer would follow up on themes encountered in previous sessions. Many definitions of semi-structured interviews suggest that the interviewer should prepare a list of questions in advance based on theory (Kallio et al., 2016), and then follow up as concepts emerge in the interview. This step was not a part of our process.

Instead, as we began to hear the students describe things that were related to intrinsic motivation, anxiety, and politeness theories, the lead interviewer (a coauthor on this paper, with extensive ethnographic interviewing experience),

started to develop questions that were related to those issues that could be used to follow up on these topics as they emerged in later interviews. For example, in order to find out about intrinsic motivation, we worked in questions like "Hey, what do you want to be when you grow up?" or "What is your favorite subject in school." The goal of these questions was to collect data that was not emerging naturally in more generic questions about the students' experiences, but where the interviewer noted that the student did not appear to have intrinsic interest in the topics being covered by Betty's Brain. These questions were then asked even in some interviews where the student's intrinsic interest had yet not emerged in the conversation. Depending on when a question emerged, it was not always possible to cover a specific topic comprehensively in these short interviews. Therefore, additional survey scales (e.g., the politeness of Mr. Davis and science anxiety scales) were added to the last day of data collection, allowing us to triangulate our interview data with other measures, as is often recommended in qualitative research guides (e.g., Leech & Onwuegbuzie, 2007).

Vignette #1: Frustration and Perceptions

Our first vignette involves the use of detector-driven classroom interviewing to study how student perceptions drive the experience of frustration during learning and in turn their choices during learning (Baker et al., 2021). For this use of DDCI, we focused most of the interviewer time and subsequent analysis on cases where the student transitioned from other affective states to frustration (priority 2), as well as when the student experienced sustained frustration for over 80 seconds (priority 1). We also investigated a small number of other transitions, including sustained confusion (over 80 seconds) (priority 3) and unresolved confusion (confusion that transitions into frustration or boredom) (priority 4).

After collecting 358 interviews, which were timestamped and labeled with triggering affective or behavioral sequences, we were able to match exactly when an interview occurred to the student's actions within the system (and the affect detector's inferences) in the 60 seconds before the interviews. We coded the interviews using grounded theory, obtaining four categories: the experience of difficulty, finding the system helpful, finding the learning content interesting, and a student reporting using strategic behaviors to support their learning. We then searched for associations between interview coding categories and affective transitions that were statistically significantly more likely than chance. We analyzed the five strongest associations in greater depth, going back to the transcripts and conducting close reading.

We found from this analysis that many of the interviews following transitions from a positive state of engaged concentration to frustration involved reports of both difficulty and strategic behavior. Students often made this transition when they could not understand the system's feedback, either from the student agent or the teacher agent (or why their answer was wrong). Studying the transcripts enabled us to identify several strategies students adopted in this situation. We also found cases where sustained frustration was associated with the experience of finding the system helpful. In this seemingly counterintuitive case, students were able to finally resolve their frustration by having the student agent take quizzes and studying the results of those quizzes; students then experienced positive attitudes towards the system as a result. As such, this analysis helped us better understand which features of the learning system were working to help keep students on track, and the strategies the students adopted when they were frustrated and found the system unhelpful.

Vignette #2: General Anxiety and Momentary Affect

We also used DDCI to examine the interplay between students' science anxiety, their affect, and their subsequent actions (Andres et al., 2022). In this work, initial detector-driven interviews produced immediate insights, which we followed up with other methods, and then analyzed the interview data in combination with other data sources to better understand the phenomena occurring around anxiety.

The initial interviews (from Vignette #1) found that a number of students in the first round of fieldwork were uncomfortable with the learning experience and appeared to be experiencing anxiety; these students talked about different strategies than students with different affective experiences.

As a result, we quickly pivoted to add an additional measure, surveying students with an adapted version of the Math Anxiety Survey (MAS) (Johnston-Wilder et al., 2014) to measure their science anxiety. We compared these survey scores to the detectors we had already used to trigger interviews, to log files of student behaviors after frustration, and to the qualitative data that emerged within the interviews that had already been conducted. We categorized students into a higher-anxiety and lower-anxiety group using a cut-off at the mean level of anxiety (producing a 48/52 split).

We then compared the anxiety survey scores to the automated detectors of frustration and found that students with higher science anxiety had different frustration patterns over time. By looking at their patterns across each day of the study, we found that students with higher anxiety appeared identical to their low-anxiety peers in the first half of each session, but experienced higher frustration in the second half of each session.

We next examined the log data of each group around periods of frustration. We found that deletion of items from a student's causal map was common following a

period of frustration, but that there were anxiety-based differences in how those deletions were enacted. Low-anxiety learners were more likely to selectively delete parts of their causal map that were related to specific errors (e.g., those that they had found from Betty's Brain's feedback on their progress). In contrast, high-anxiety learners cleared a larger number of links in a seemingly haphazard manner, a pattern that corresponded with reported behavior and observations by interviewers. This investigation of deletion patterns was initially driven by the qualitative observations of the *in situ* interviewers.

We subsequently sought to determine whether we could better understand these patterns by conducting a quantitative analysis connecting the interview transcripts, the automated measurements of frustration, and the survey measurements of anxiety. For this analysis, we examined all of the interviews that were conducted within the 80 seconds after frustration being inferred by the automated detector (28 of the total 594 interviews). We found that high anxiety students were more likely to discuss deleting significant parts of their concept map, but also discussed their motivations for doing so. Specifically, these students consistently reported being so overwhelmed and confused that this more drastic approach was warranted. In other words, high-anxiety students appear less able to regulate their learning process, fixating instead on the source of their frustration instead of activities or resources that could aid in resolving them. This may have caused them to either overlook or deliberately skip activities that might be essential for their learning (Ashcraft, 2002; Browning et al., 2015), possibly because they did not recognize the opportunities provided to them.

Vignette #3: Verbal Metacognitive Expressions

In this Vignette, we discuss secondary data analysis of the interview transcripts from Vignette #1, conducted with the goal of understanding students' verbal metacognitive expressions (Bosch et al., 2021). To conduct this analysis, we transcribed the text of the interviews and then applied a natural language processing tool that can identify metacognitive expressions in text (Bosch et al., 2019), counting phrases beginning with a first-person pronoun and ending with a metacognitive indicator word, such as "considered" or "expected". Metacognitive expressions correlated negatively with occurrences of unresolved confusion (i.e., when students' confusion eventually turned to boredom or frustration), as measured by the real-time affect detectors that we also used to drive *in situ* interviews. Metacognitive expressions also correlated positively with behavioral indicators of metacognition, further suggesting that the combination of affect detectors and natural language processing can help qualitative researchers to investigate interesting points in time where students are working through problems metacognitively. We also found that the interviews themselves may have promoted metacognition by encouraging students to think through issues they were facing as they explained them. Furthermore, we found evidence that participating in interviews was associated with better learning gains for students who demonstrated more metacognitive reasoning.

Vignette #4: Re-design of Tutor Communication

One somewhat surprising issue that emerged from our in situ interviews for vignette #1 was that students were having negative reactions to Mr. Davis, the virtual mentor who provides advice to students as they progress through Betty's Brain. As we interviewed students during the vignette #1 data collection, it was common for

students to report that Mr. Davis was rude and unhelpful, and some were distressed by their interactions with him, with one student for example referring to him as "messed up", another stating "I know for a fact that he doesn't actually care", and a third stating "he goes crazy".

Research findings on politeness in virtual tutors have been mixed, which is perhaps not surprising given the many functions that politeness strategies might have in tutoring sessions run by humans (e.g., review by Ogan et al., 2012). Several studies have shown that, in some contexts, students respond well to so-called "rude tutors" who are designed to offer sarcastic responses even to struggling students (e.g., Graesser, 2011; Ogan et al., 2012), but other studies have shown that students respond better to more polite approaches (Graesser 2011; Tynan, 2005), and some researchers have found that learning decreases when students perceive their tutors as irritating or offensive (De Angeli & Brahnam, 2008).

In the interviews, students reported that they were upset by their interactions with Mr. Davis' interactions. Students discussed a range of complaints about the system, from the system interrupting the student too often, or that he would answer "I don't know" to their questions. However, many students discussed how Mr. Davis' statements began with "Hmph" or that Mr. Davis' would say things like "good thing that I'm not your teacher."

We then began asking what changes would make Mr. Davis' appear more polite. Students repeatedly suggested changing "Hmph" to "Hmm," removing some of the other rude phrases, sounding more attentive and warm, and offering more help when they were struggling, to better match the mentor role.

We modified Betty's Brain according to these suggestions, changing "Hmph" to "Hmm" and adding new scripts to make Mr. Davis' better match the students' expectations for mentorship conversations. These included both scripts that

provided encouragement (a very simple change) and scripts that offered additional help via strategic hints (a slightly more complicated strategy).

Table 1. Modifications to Mr. Davis' script

We then conducted a second round of DDCI research on a new unit (with the same students), also administering a quantitative instrument on students' perceptions of system helpfulness and difficulty. These changes led to students finding Mr. Davis more helpful, and also led to the students finding the system less difficult in general. Improved perception of Mr. Davis was associated with better learning outcomes for these students. Students reported during interviews, when asked if anything had changed from the last version, that Mr. Davis was now more helpful: "It helps when he, like, one time he told me to delete a link and I did and it really helped… he said the specific information, said to delete the one from heat inside the body I think to heat loss. And that actually really helped my quiz results, here.", "Yeah, and he helped, he helped definitely. He's more helpful this unit, definitely."

Students also reported that Mr. Davis had a more positive personality: "He's like, how are you feeling?, which was nice,", "Mr. Davis was a lot meaner last time, and he's nicer now." although students still complained that Mr. Davis intervened too often or that he did not tell them the answer.

These results demonstrate that even minor changes at the pragmatics level of conversation (i.e., changes that effect the perception of politeness and intent) may impact students' interactions with a learning system. It also demonstrates that—in addition to helping to explore how politeness may interact with other parts of the design of computer-based learning environments—the rapid collection of qualitative data that DDCI facilitates can expedite real-world design iteration and

improvement, in identifying problems in design, identifying possible solutions in partnership with learners, and evaluating the results.

Comparison/Contrast to Related Methodologies

As the vignettes above illustrate, DDCI can be used within a range of mixed methods studies. DDCI is itself a mixed methods approach, using quantitative methods in order to select cases for qualitative research and inform the researcher's decisions around what questions to ask. In contrast to some other mixed methods approaches, DDCI does not use one type of method to explain the other type of method's results, and does not use the two types of methods convergently (even if the overall program of research involves triangulation between methods). Instead, the goal of DDCI is to target and focus qualitative research using quantitative methods—to address the needle in a haystack problem and enable real-time *in situ* research on intermittent, occasional, or rare phenomena. In the following subsections, we compare and contrast three related types of mixed methods research to DDCI, to better elucidate what is different about DDCI compared to past work.

Computational Grounded Theory

We first compare DDCI to computational grounded theory (Nelson 2020), which shares with DDCI a focus on using computational methods to support qualitative work. Computational grounded theory consists of three steps: 1) unsupervised learning to extract candidate text patterns automatically, 2) human expert interpretation of the data related to step 1 patterns, and 3) automating measurement of interesting patterns from step 2 throughout an entire text corpus. Steps 1 and 2 in particular are related to DDCI, in that data science methods (step 1) are applied to determine where to focus qualitative efforts (step 2). In step 2 of computational

grounded theory, a qualitative analyst might perform deep reading of text samples that contain the patterns identified in step 1 to distinguish patterns that capture something meaningful—and perhaps even unexpected—from spurious patterns. Similarly, a qualitative analyst using our proposed methodology might apply a method like thematic analysis (Clarke & Braun, 2017) to uncover the meaning of patterns identified via data science methods.

In DDCI, data science methods serve to focus the efforts of qualitative research, especially in situations where analysis of an entire dataset is intractable. The findings are ultimately qualitative in nature. The goals of computational grounded theory, however, are somewhat more quantitative. Computational grounded theory results in scalable, quantitative measures of constructs identified in text, thereby leveraging the power of computational approaches to work at large scale. Qualitative expertise ensures that quantitative measures of these constructs are interpretable and meaningful. Thus, computational grounded theory offers scalability, while our proposed approach offers depth.

Data-Driven Retrospective Interviewing

A second approach related to the approach presented in our paper is called Data-Driven Retrospective Interviewing (DDRI), which was used by El-Nasr et al. (2015) to evaluate a game. In that study, metrics and visualizations of gameplay were distilled from data logs of gameplay, including metrics such as the type of quests completed and time spent, and visualizations such as bar graphs of how many times each player engaged in specific activities within the game. These metrics and visualizations were then used to decide what topics to focus on in a set of retrospective interviews, and were used as artifacts to drive discussion as each player was interviewed. While DDRI used data to inform the content of the

interviews, it differs from the DDCI method in that DDRI does not use data (*in situ* or otherwise) to select which individuals to interview. Within El-Nasr et al.'s (2015) method, data was used in a more intensive fashion during interviews than in DDCI, as interviewers showed data directly to interview subjects and discussed it with them.

Using Quantitative Analysis to Select Cases for Qualitative Analysis

Leary et al. (2021) also propose an approach related to the approach proposed in this paper. Their approach involved first conducting data mining on log data of teachers' usage of a curriculum planning tool. They found that teachers varied in terms of two dimensions of use: frequency and variability. They then looked at existing qualitative data from the teachers with high variability in order to understand the implications of that variability. They had already conducted interviews and observations of teachers, but without the use of any method for focusing the qualitative data collection; instead, interviews were conducted exhaustively and observations were conducted according to convenience. In other words, quantitative data analysis drove the selection of cases for qualitative analysis but not the selection of cases for qualitative data collection. As such, their approach focused researcher time during one phase of the qualitative research process, but not the other phase.

Potential Applications and Opportunities

In this paper, we have discussed four examples of the use of detector-driven classroom interviewing, showing how this approach can be used to study why students become frustrated and how they respond, how anxiety influences how students respond to their own frustration, how metacognition interacts with affect,

and how to improve the design of an adaptive learning system. These examples, all conducted within the same year, are a first illustration of the breadth of research this approach affords, as well as the relatively high speed with which it enables qualitative research in education.

There are a range of opportunities in research and practice for detector-driven classroom interviewing. We group these roughly into three categories: 1) research opportunities, 2) design improvement opportunities, 3) pedagogical opportunities for teachers.

Research Opportunities of Detector-Driven Classroom Interviewing

This article has demonstrated a small number of the cases where DDCI was useful, all taken from the context of the Betty's Brain learning platform. However, there are many more open problems and challenges in research around self-regulated learning, metacognition, and affect that detector-driven classroom interviewing could shed light on.

The method's success with studying the moment-to-moment affect and self-regulatory behaviors associated with anxiety demonstrates that DDCI can support research on coarser-grained manifestations of affect and mood and their development over time. Affective research has struggled to represent and study affective manifestations with durations longer than the tens of seconds and shorter than months. This challenge can be seen, for example, in the split between state anxiety and trait anxiety (Endler & Kocovski, 2001). While theoretical models and empirical results are clear about the split between these two forms of anxiety, perhaps there is a temporally in-between "mood" anxiety which manifests over the course of days. This mood anxiety could be identified at first by repeated questionnaires and later by behavioral manifestations in learning activities, and then

studied in depth via interviews targeted to when mood anxiety increases or decreases. Indeed, despite evidence for some forms of affect manifesting over much longer durations than others (Verduyn & Lavrijsen, 2015) and some longer-duration emotions co-occurring with shorter-term affective states (Author, year; Author, year; Dillon et al., 2016), the interplay between brief affect and longer-term moods and emotions during learning has not been thoroughly studied.

Detector-driven classroom interviews may also help us understand the emerging evidence that affective states can co-occur (Dillon et al., 2016). By detecting the co-occurrence of affective states (affect detectors have long identified co-occurring affect but the phenomenon has been treated in practice as uncertainty or detector error), we can trigger interviews to understand the phenomenological experience of being both bored and frustrated at the same time, for instance.

Overall, this method provides a way to drill deeper into the details of affective experience. One can envision several other ways to use this method to study key affective experiences. For example, it may be useful to compare the in-the-moment affective experiences (e.g., confusion and frustration) between students who are persisting productively and students who are persisting unproductively. Doing so could lead to better methods for scaffolding students who are persisting unproductively, helping them to develop better strategies to make their persistence productive. Similarly, using DDCI to study the experiences of students who frequently experience boredom while using a learning system could help designers to better understand the different factors leading to boredom and how to support different learners (e.g., students who need more challenging material vs. students whose frustration turns to boredom, suggesting that they need more scaffolding).

A second future research application of these methods is to better understand the phenomenological experience of disengagement. Despite decades of research on disengagement during learning, only limited research has used methods such as interviews to gain students' own perspectives on their experience of disengagement during online learning (but see Aagaard, 2015; Schofield, 1995; Xia et al., 2020). DDCI could be used to more deeply study the phenomenological experience of becoming disengaged, and what occurs during disengagement. What is the emotional experience and reasoning that occurs when a learner is gaming the system to complete content without learning (Baker et al., 2004; Xia et al., 2020) or responding carelessly (Clements, 1982)? Relatively more work has occurred in understanding the experience of mind wandering (Stawarczyk et al., 2011), using experience sampling methods, but this work has not yet drilled down selectively into specific situations where mind wandering occurs. More focused methods for studying mind wandering may help us to understand mind wandering in greater depth, exploring phenomena such as the positive role that mind wandering sometimes plays in successful work performance (e.g., Baird et al., 2012) and helping us design ways to help learners better regulate and even harness their mind wandering.

Design Improvement Opportunities in Detector-Driven Classroom Interviewing

Beyond research, detector-driven classroom interviewing provides many opportunities for improving design as well. The contemporary design of learning activities often involves design experiments/design research in classrooms. However, the qualitative component of classroom design experiments is currently

time-consuming and suffers from the needle in a haystack problem discussed earlier.

Much of the process of refining design comes from identifying and fixing critical moments in the learning experience: a moment where a learner struggles and fails to resolve their struggle (Nawaz et al., 2018), a moment where a learner's frustration transitions to boredom (Andres et al., 2019), or a moment where a learner disengages completely (Botelho et al., 2019). Currently, if a researcher is not lucky enough to be watching a student when one of these events occurs, finding these critical events requires retrospective log data analysis or discussions with students long after the event has passed.

Log data analysis generally is unable to provide a deeper understanding of why these shifts occur—what changes in a student's perception, attitudes, or cognition produce these shifts. Hearing about a shift retrospectively may miss key cognition surrounding the event—a learner may no longer remember exactly what happened and why (van Someren et al., 1994). And only a small proportion of events will be captured if the classroom researcher is dependent on hearing an exclamation or watching the right student at the right time. By capturing critical moments and understanding what happened during and right before these critical events, we can re-design learning activities to minimize disruption to successful student learning. We can identify the factors that most increase student difficulty and better scaffolding them (Rittle-Johnson & Koedinger, 2005), clarify confusing interface elements (Vermeeren et al., 2007), and re-design messages and agent-based interactions that annoy or bore students, as discussed in vignette #4 above (Ocumpaugh et al., 2021).

The understanding that detector-driven classroom interviewing brings can also be a useful tool in co-design and participatory design efforts, particularly for

developers of learning systems designing technologies that will be used by historically marginalized communities that they are not themselves members of. As noted above, targeted qualitative data collection facilitates discussion at critical moments during an experience. These discussions can be opportunities for students to provide in-the-moment and unfiltered feedback that can influence the design, or even to engage in rapid mini-design sessions where students propose solutions to the flaws in the system they are experiencing. By attempting to understand a problem immediately when it occurs, we can focus design efforts on the exact elements of the learning system associated with negative learner experiences. By conducting this process with members of the communities we are designing for, we can more fully include them in the design process and center the design in their experiences. This approach recognizes the value of students' perspective on their learning, and promotes a sense of ownership in the development process. DDCI allows students to communicate in their own way at key moments in their learning experience. Rather than rely on predefined options (e.g., multiple choice), students can express themselves (and their identity) in the way that is most natural for them, helping to enable fairer and more meaningful participation in design decisions (Costanza-Chock, 2018). As noted by Costanza-Chock (2018, p. 10), "the tacit and experiential knowledge of community members is sure to produce ideas, approaches and innovations that a non-member of the community would be very unlikely to come up with." By obtaining student perspectives at the key moments where a design fails, we can obtain information that might not be available in an out-of-context design session or focus group (Dumas et al., 1999).

Pedagogical Opportunities for Teachers

A third key area of opportunity comes not from the interviewing, per se, but from the infrastructure developed for detector-driven classroom interviewing. Such an infrastructure provides an opportunity for teachers, which can be used to focus teachers time more effectively. The use of educational technology in classrooms creates opportunities for teachers to focus their time on sub-groups of students while other students make progress on their own (Schofield, 1995), a practice that has developed even without the use of detectors.

For example, practices such as proactive remediation, where a teacher obtains information from a learning platform on a specific student's progress and reaches out to them to offer assistance, have been encouraged by existing systems that simply provide real-time data on student correctness (Miller et al., 2015). However, existing dashboards and reports for teachers are typically limited to summarizing student performance and knowledge (Jivet et al., 2017), both limiting the range of situations teachers can address to a student struggling with a specific topic, and requiring teachers to develop skill at rapidly interpreting those reports. Increasingly, teachers are interested in obtaining richer, real-time information (Holstein et al., 2019), and the infrastructure used in DDCI could be leveraged to inform teachers about more than just whether a student is obtaining incorrect answers. It could be used to notify teachers about a range of events: a disengaged student (to be checked in on), a student applying a positive self-regulated learning strategy (to be praised and perhaps used as a model for other students), or a student making progress but using an ineffective strategy (to be scaffolded in working more effectively). Teachers could use the interview-triggering part of the infrastructure—not to conduct interviews for research purposes—but to spark brief, in-the-moment timely conversations with struggling students. These conversations could help the

teacher understand why a specific student is struggling with something, and in particular to understand if that student is struggling with something that seems like it should be easy—helping teachers discover their own expert blind spots (Nathan & Petrosino, 2003).

Conclusions

In this paper, we have described our work using detector-driven classroom interviewing (DDCI), a form of mixed methods research that uses quantitative methods (machine learned automated detectors of student states and strategies during classroom learning) in service of qualitative methods (interviewing). The goal of DDCI is to conduct interviews in real-time, *in situ*, as learners engage in an educational activity. Without such an approach, it is difficult for a researcher to focus their time and attention on more important events. Some interesting and important events may only occur rarely, intermittently, and briefly (i.e., the needle-in-a-haystack problem). An interviewer must visit many classrooms to capture these events. With DDCI, the interviewer is notified when an event of interest has recently occurred, and can quickly interview the student.

We provide a case study of the application of DDCI in the context of studying phenomena surrounding student affect and self-regulated learning in the Betty's Brain Open Ended Learning Environment. We present the Quick Red Fox (QRF) app used to notify researchers about events of interest, and briefly discuss the automated detectors used to identify the events. We then discuss four vignettes of how DDCI supported mixed methods research that was published, within this context.

Our article discusses the differences between DDCI and related recent methodologies, and also discusses the potential opportunities of DDCI for

promoting educational research, educational design, and teaching. It is our view that DDCI, by helping to focus the scarce resources of researcher human attention and time, can facilitate qualitative research and drive down the amount of effort needed to conduct the types of research needed to help us come closer to genuinely understanding learners and how they can be better supported by learning technologies.

Acknowledgments

REDACTED for review.

References

Aagaard, J. (2015). Drawn to distraction: A qualitative study of off-task use of educational technology. Computers & Education, 87, 90-97.

Aldiabat, K. M., & Le Navenec, C. L. (2018). Data saturation: The mysterious step in grounded theory methodology. The qualitative report, 23(1), 245-261.

Andres, J.M.A.L., Ocumpaugh, J., Baker, R., Slater, S., Paquette, S., Jiang, Y., Bosch, N., Munshi, A., Moore, A., Biswas, G. (2019). Affect sequences and learning in Betty's Brain. Proceedings of the 9th International Learning Analytics and Knowledge Conference, 383-390.

Andres. J.M.A.L., Hutt, S., Ocumpaugh, J., Baker, R.S., Nasiar, N., Porter, C. (2022). How anxiety affects affect: A quantitative ethnographic investigation using affect detectors and data-targeted interviews. In Advances in Quantitative Ethnography: Third International Conference, ICQE 2021, Virtual Event, November 6–11, 2021, Proceedings 3 (pp. 268-283). Springer International Publishing.

Ashcraft, M. H. (2002). Math anxiety: Personal, educational, and cognitive consequences. Current Directions in Psychological Science, 11(5), 181-185. https://doi.org/10.1111/1467-8721.00196

Azevedo, R., & Gašević, D. (2019). Analyzing multimodal multichannel data about self-regulated learning with advanced learning technologies: Issues and challenges. Computers in Human Behavior, 96, 207-210. https://doi.org/10.1016/j.chb.2019.03.025

Azevedo, R., Moos, D. C., Johnson, A. M., & Chauncey, A. D. (2010). Measuring cognitive and metacognitive regulatory processes during hypermedia learning: Issues and challenges. Educational Psychologist, 45(4), 210-223.

Baird, B., Smallwood, J., Mrazek, M. D., Kam, J. W., Franklin, M. S., & Schooler, J. W. (2012). Inspired by distraction: Mind wandering facilitates creative incubation. Psychological Science, 23(10), 1117-1122. https://doi.org/10.1177/0956797612446024

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

Baker, R.S., Corbett, A.T., Koedinger, K.R., Wagner, A.Z. (2004). Off-task behavior in the cognitive tutor classroom: When students "game the system". Proceedings of ACM CHI 2004: Computer-Human Interaction, 383-390.

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. (2021). Affect-targeted interviews for understanding student frustration. Proceedings of the International Conference on Artificial Intelligence and Education.

Baker, R.S., Ocumpaugh, J.L., Andres, J.M.A.L. (2020). BROMP quantitative field observations: A review. In R. Feldman (Ed.) Learning Science: Theory, Research, and Practice, pp. 127-156. New York, NY: McGraw-Hill.

Baker, R.S.J.d. & Rossi, L.M. (2013). Assessing the disengaged behavior of learners. In Sottilare, R., Graesser, A., Hu, X., & Holden, H. (Eds.) Design Recommendations for Intelligent Tutoring Systems – Volume 1 – Learner Modeling. U.S. Army Research Lab, Orlando, FL, pp. 155-166, 2013.

Barrett, L. F. (2017). The theory of constructed emotion: an active inference account of interoception and categorization. Social cognitive and affective neuroscience, 12(1), 1-23.

Barriball, K. L., & While, A. (1994). Collecting data using a semi-structured interview: a discussion paper. Journal of Advanced Nursing-Institutional Subscription, 19(2), 328-335.

Biswas, G., Baker, R. S., & Paquette, L. (2017). Data mining methods for assessing self-regulated learning. In Handbook of Self-Regulation of Learning and Performance (pp. 388-403). Routledge.

Bosch, N., & D'Mello, S. (2014). It takes two: Momentary co-occurrence of affective states during computerized learning. In International Conference on Intelligent Tutoring Systems (pp. 638-639). Springer, Cham.

Bosch, N., Huang, E., Angrave, L., & Perry, M. (2019). Modeling improvement for underrepresented minorities in online STEM education. In Proceedings of the 27th ACM Conference on User Modeling, Adaptation and Personalization (pp. 327-335).

Bosch, N., Zhang, Y., Paquette, L., Baker, R. S., Ocumpaugh, J., & Biswas, G. (2021). Students' verbalized metacognition during computerized learning. Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (CHI '21), 680:1-680:12. https://doi.org/10.1145/3411764.3445809

Botelho, A. F., Varatharaj, A., Inwegen, E. G. V., & Heffernan, N. T. (2019). Refusing to try: Characterizing early stopout on student assignments. In Proceedings of the 9th International Conference on Learning Analytics & Knowledge, 391-400. https://doi.org/10.1145/3303772.3303806

Botelho, A.F., Baker, R., Ocumpaugh, J., Heffernan, N. (2018). Studying affect dynamics and chronometry using sensor-free detectors. Proceedings of the 11th International Conference on Educational Data Mining, 157-166.

Briggs, C. L. (1986). Learning how to ask: A sociolinguistic appraisal of the role of the interview in social science research. Cambridge university press.

Browning, M., Behrens, T. E., Jocham, G., O'Reilly, J. X., & Bishop, S. J. (2015). Anxious individuals have difficulty learning the causal statistics of aversive environments. Nature Neuroscience, 18(4), 590. https://doi.org/10.1038/nn.3961

Clarke, V., & Braun, V. (2017). Thematic analysis. The Journal of Positive Psychology, 12(3), 297–298. https://doi.org/10.1080/17439760.2016.1262613

Clements, M. (1982). Careless errors made by sixth-grade children on written mathematical tasks. Journal for Research in Mathematics Education, 13(2), 136-144. https://doi.org/10.1016/j.cedpsych.2019.01.007

Costanza-Chock, S. (2018). Design Justice, AI, and Escape from the Matrix of Domination. Journal of Design and Science. https://doi.org/10.21428/96c8d426

De Angeli, A., & Brahnam, S. (2008). "I hate you! Disinhibition with virtual partners." Interacting with Computers 20(3), 302–10. https://doi.org/10.1016/j.intcom.2008.02.004

DeFalco, J.A., Rowe, J.P., Paquette, L., Georgoulas-Sherry, V., Brawner, K., Mott, B.W., Baker, R.S., Lester, J.C. (2018). Detecting and addressing frustration in a serious game for military training. International Journal of Artificial Intelligence and Education, 28 (2), 152-193. https://doi.org/10.1007/s40593-017-0152-1

Dillon, J., Bosch, N., Chetlur, M., Wanigasekara, N., Ambrose, G. A., Sengupta, B., & D'Mello, S. K. (2016). Student emotion, co-occurrence, and dropout in a MOOC context. Proceedings of the International Conference on Educational Data Mining.

D'Mello, S. (2013). A selective meta-analysis on the relative incidence of discrete affective states during learning with technology. Journal of Educational Psychology, 105(4), 1082. https://doi.org/10.1037/a0032674

D'Mello, S., & Graesser, A. (2011). The half-life of cognitive-affective states during complex learning. Cognition & Emotion, 25(7), 1299-1308. https://doi.org/10.1080/02699931.2011.613668

D'Mello, S., Lehman, B., Sullins, J., Daigle, R., Combs, R., Vogt, K., Perkins, L., & Graesser, A. (2010). A time for emoting: When affect-sensitivity is and isn't effective at promoting deep learning. In International Conference on Intelligent Tutoring Systems, 245-254. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-13388-6_29

Dumas, J. S., Dumas, J. S., & Redish, J. (1999). A practical guide to usability testing. Intellect Books.

El-Nasr, M. S., Durga, S., Shiyko, M., & Sceppa, C. (2015). Data-driven retrospective interviewing (DDRI): a proposed methodology for formative evaluation of pervasive games. Entertainment Computing, 11, 1-19. https://doi.org/10.1016/j.entcom.2015.07.002

Eder, D. & Fingerson, L. (2002) Interviewing Children and Adolescents in Gubrium, J.K. & Holstein, J.A., Handbook of Interview Research. London: Sage.

Emmel, N. (2013). Purposeful sampling. Sampling and choosing cases in qualitative research: A realist approach, 33-45. https://dx.doi.org/10.4135/9781473913882.n3

Endler, N. S., & Kocovski, N. L. (2001). State and trait anxiety revisited. Journal of Anxiety Disorders, 15(3), 231-245. https://doi.org/10.1016/S0887-6185(01)00060-3

Erbas, A. K., & Okur, S. (2012). Researching students' strategies, episodes, and metacognitions in mathematical problem solving. Quality & Quantity, 46, 89-102.

Goffman, E. (1959). The presentation of self in everyday life. New York: Doubleday.

Graesser, A. C. (2011). Learning, thinking, and emoting with discourse technologies. American Psychologist, 66(8), 746. https://doi.org/10.1037/a0024974

Hershkovitz, A., Baker, R. S., Gobert, J., Wixon, M., & Pedro, M. S. (2013). Discovery with models: A case study on carelessness in computer-based science inquiry. American Behavioral Scientist, 57(10), 1480-1499. https://doi.org/10.1177/0002764213479365

Hoeber, O., Hoeber, L., El Meseery, M., Odoh, K., & Gopi, R. (2016). Visual Twitter analytics (Vista): Temporally changing sentiment and the discovery of emergent themes within sport event tweets. Online Information Review, 40(1), 25-41. https://doi.org/10.1108/OIR-02-2015-0067

Holstein, K., Wortman Vaughan, J., Daumé III, H., Dudik, M., & Wallach, H. (2019, May). Improving fairness in machine learning systems: What do industry practitioners need?. In Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, 1-16. https://doi.org/10.1145/3290605.3300830

Hunting, R. P. (1997). Clinical interview methods in mathematics education research and practice. The Journal of Mathematical Behavior, 16(2), 145-165.

Hutt, S., Baker, R. S., Ocumpaugh, J., Munshi, A., Andres, J. M. A. L., Karumbaiah, S., Slater S., Biswas G., Paquette L., Bosch, N. & van Velsen, M. (2022). Quick red fox: an app supporting a new paradigm in qualitative research on AIED for STEM. Artificial Intelligence in STEM Education: The Paradigmatic Shifts in Research, Education, and Technology, 319-332.

Hutt, S., Grafsgaard, J. F., & D'Mello, S. K. (2019). Time to scale: Generalizable affect detection for tens of thousands of students across an entire school year. In Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (pp. 1-14).

Jerolmack, C., & Khan, S. (2014). Talk is cheap: Ethnography and the attitudinal fallacy. Sociological methods & research, 43(2), 178-209.

Jiang, Y., Paquette, L., Baker, R.S., Clarke-Midura, J. (2015) Comparing Novice and Experienced Students in Virtual Performance Assessments. Proceedings of the 8th International Conference on Educational Data Mining, 136-143.

Jivet, I., Scheffel, M., Drachsler, H., & Specht, M. (2017). Awareness is not enough: Pitfalls of learning analytics dashboards in the educational practice. In European Conference on Technology Enhanced Learning, 82-96. Springer, Cham. https://doi.org/10.1007/978-3-319-66610-5_7

Johnston-Wilder, S., Brindley, J., & Dent, P. (2014). A survey of mathematics anxiety and mathematical resilience among existing apprentices. London: The Gatsby Foundation.

Kallio, H., Pietilä, A. M., Johnson, M., & Kangasniemi, M. (2016). Systematic methodological review: developing a framework for a qualitative semi-structured interview guide. Journal of advanced nursing, 72(12), 2954-2965.

Knapp, N. F. (1997). Interviewing Joshua: On the importance of leaving room for serendipity. Qualitative Inquiry, 3(3), 326-342. https://doi.org/10.1177/107780049700300305

Kvale, S., & Brinkmann, S. (2009). Interviews: Learning the craft of qualitative research interviewing. Sage.

Labov, W. (1972). Some principles of linguistic methodology. Language in society, 1(1), 97-120.

Leary, H., Lee, V. R., & Recker, M. (2021). It's more than just technology adoption: Understanding variations in teachers' use of an online planning tool. TechTrends, 65(3), 269-277. https://doi.org/10.1007/s11528-020-00576-3

Leech, N. L., & Onwuegbuzie, A. J. (2007). An array of qualitative data analysis tools: A call for data analysis triangulation. School psychology quarterly, 22(4), 557.

Leelawong, K., & Biswas, G. (2008). Designing learning by teaching agents: The Betty's Brain system. International Journal of Artificial Intelligence in Education, 18(3), 181-208.

Lindquist, K. A., Barrett, L. F., Bliss-Moreau, E., & Russell, J. A. (2006). Language and the perception of emotion. Emotion, 6(1), 125.

Luo, G. (2015). MLBCD: a machine learning tool for big clinical data. Health Information Science and Systems, 3(1), 1-19. https://doi.org/10.1186/s13755-015-0011-0

Miller, W.L., Baker, R., Labrum, M., Petsche, K., Liu, Y-H., Wagner, A. (2015) Automated detection of proactive remediation by teachers in Reasoning Mind classrooms. Proceedings of the 5th International Learning Analytics and Knowledge Conference, 290-294. https://doi.org/10.1145/2723576.2723607

Munshi, A., Biswas, G., Baker, R., Ocumpaugh, J., Hutt, S., & Paquette, L. (2023). Analysing adaptive scaffolds that help students develop self-regulated learning behaviours. Journal of Computer Assisted Learning, 39(2), 351-368.

Munshi, A., Rajendran, R., Ocumpaugh, J., Biswas, G., Baker, R. S., & Paquette, L. (2018, July). Modeling learners' cognitive and affective states to scaffold SRL in open-ended learning environments. *In Proceedings of the 26th Conference on User modeling, Adaptation and Personalization* (pp. 131-138).

Nathan, M. J., & Petrosino, A. (2003). Expert blind spot among preservice teachers. American Educational Research Journal, 40(4), 905-928. https://doi.org/10.3102/00028312040004905

Nawaz, S., Kennedy, G., Bailey, J., Mead, C., & Horodyskyj, L. (2018). Struggle town? Developing profiles of student confusion in simulation-based learning environments. In 35th International Conference on Innovation, Practice and Research in the Use of Educational Technologies in Tertiary Education, ASCILITE, 224-233.

Nelson, L. K. (2020). Computational grounded theory: A methodological framework. Sociological Methods & Research, 49(1), 3-42. https://doi.org/10.1177/0049124117729703

Ocumpaugh, Jaclyn, Stephen Hutt, J. M. A. L. Andres, Ryan S. Baker, Gautam Biswas, Nigel Bosch, Luc Paquette, and Anabil Munshi. "Using qualitative data from targeted interviews to inform rapid AIED development." In Proceedings of the 29th international conference on computers in education, pp. 69-74. 2021.

Ogan, A., Finkelstein, S., Mayfield, E., D'Adamo, C., Matsuda, N., & Cassell, J. (2012). "Oh dear Stacy!" Social interaction, elaboration, and learning with teachable agents. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, 39-48. https://doi.org/10.1145/2207676.2207684

Paquette, L., Grant, T., Zhang, Y., Biswas, G., & Baker, R. (2021). Using epistemic networks to analyze self-regulated learning in an open-ended problem-solving environment. In International Conference on Quantitative Ethnography (pp. 185-201). Springer, Cham.

Patton, M. Q. (2002). Qualitative Research and Evaluation Methods (3rd ed.). Thousand Oaks, CA: Sage.

Ravitch, S. M., & Carl, N. M. (2019). Qualitative research: Bridging the conceptual, theoretical, and methodological. Sage Publications.

Rittle-Johnson, B., & Koedinger, K. R. (2005). Designing knowledge scaffolds to support mathematical problem solving. Cognition and Instruction, 23(3), 313-349. https://doi.org/10.1207/s1532690xci2303_1

Roll, I., Aleven, V., McLaren, B. M., & Koedinger, K. R. (2011). Improving students' help-seeking skills using metacognitive feedback in an intelligent tutoring system. Learning and instruction, 21(2), 267-280. https://doi.org/10.1016/j.learninstruc.2010.07.004

Saldana, J. (2011). Fundamentals of qualitative research. OUP USA.

Schofield, J. W. (1995). Computers and classroom culture. Cambridge University Press.

Schooler, J. W., Ohlsson, S., & Brooks, K. (1993). Thoughts beyond words: When language overshadows insight. Journal of Experimental Psychology: General, 122(2), 166. https://doi.org/10.1037/0096-3445.122.2.166

Seidman, I. (2006). Interviewing as qualitative research: A guide for researchers in education and the social sciences. Teachers college press.

Spradley, J. P. (2016). The Ethnographic Interview. Waveland Press.

Stawarczyk, D., Majerus, S., Maj, M., Van der Linden, M., & D'Argembeau, A. (2011). Mind-wandering: Phenomenology and function as assessed with a novel experience sampling method. Acta Psychologica, 136(3), 370-381. https://doi.org/10.1016/j.actpsy.2011.01.002

Strauss, A., & Corbin, J. (1990). Basics of qualitative research. Sage publications.

Tynan, R. (2005). The effects of threat sensitivity and face giving on dyadic psychological safety and upward communication 1. Journal of Applied Social Psychology, 35(2), 223-247. https://doi.org/10.1111/j.1559-1816.2005.tb02119.x

Van Someren, M. W., Barnard, Y. F., & Sandberg, J. A. (1994). The think aloud method: A practical approach to modelling cognitive. London: Academic Press.

Verduyn, P., & Lavrijsen, S. (2015). Which emotions last longest and why: The role of event importance and rumination. Motivation and Emotion, 39(1), 119-127. https://doi.org/10.1007/s11031-014-9445-y

Vermeeren, A. P. O. S., Bekker, M. M., Kesteren, I. V., & Ridder, H. D. (2007). Experiences with structured interviewing of children during usability tests. In Proceedings of HCI 2007 The 21st British HCI Group Annual Conference University of Lancaster, UK 21, 1-9. https://doi.org/10.14236/ewic/HCI2007.14

Ward, M.D. (1981) The Observer Effect in Classroom Visitation. Unpublished doctoral dissertation, Brigham Young University.

Wengraf, T. (2001). Qualitative research interviewing: Biographic narrative and semi-structured methods. sage.

Xia, M., Asano, Y., Williams, J. J., Qu, H., & Ma, X. (2020). Using information visualization to promote students' reflection on "gaming the system" in online learning. In Proceedings of the Seventh ACM Conference on Learning@ Scale, 37-49. https://doi.org/10.1145/3386527.3405924

Appendices

Figure 1. Screenshot of viewing quiz results and checking the chain of links Betty used to answer a quiz question.

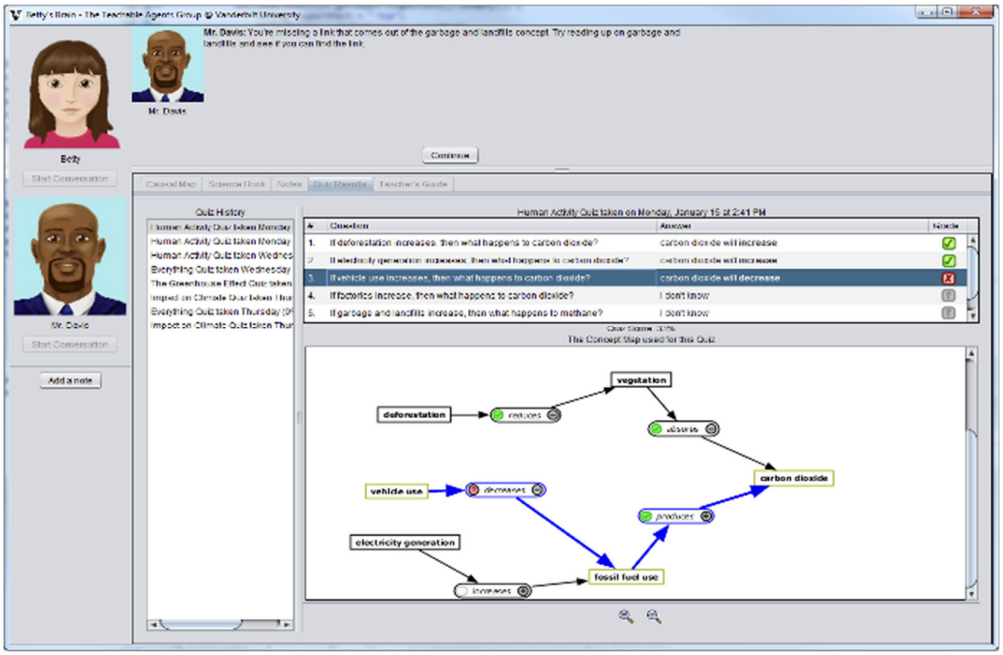

Figure 2. The Quick Red Fox app for classroom interviewing.

Table 1. Modifications to SIMULATED-TEACHER-NAME' script.

Modification Type	Example
Representational change for politeness	Hmph → Hmm
Additional feedback with encouragement	"Looks like you are doing a good job teaching correct causal links to Betty! Make sure that you check her progress from time to time by asking her to take a quiz."
	"Wow! I think I have some correct links on the map. This is fun! Thanks, A."
Additional hints/guidance	"Hey, from the quiz results, it looks like Betty may have some incorrect links on her map. You can mark those links as 'could be wrong' in your map. Do you want to know more about marking links as 'could be wrong'?"
	"You are missing a link that comes out of 'heat loss'. Try reading up on Page 'Response 1: Skin Contraction' and see if you can find the link."
	"From the quiz, it seems you may have an incorrect shortcut link on your map. Do you want to know more about shortcut links?"

Figure 3. Example of ENA network visualization taken from the analysis of the relationship between coherent actions in Betty's Brain (Paquette et al. 2021) -- relationships between behaviors for high-performing learners are shown in blue, whereas those relationships are shown in red for low-performing learners.