SITUATING BIG DATA:

Assessing Game-Based STEM Learning in Context

Summative Evaluation Report

Prepared by:

Rockman et al
Research & Evaluation

Project Credits:
This study is supported by the National Science Foundation's Division of Research on Learning (Award No. 1418352). Any opinions, findings, conclusions, or recommendations expressed in this report are those of the evaluation team and do not necessarily reflect the views of the National Science Foundation.

For permission requests, write to:
Rockman et al
201 Mission Street, Suite 1320
San Francisco, CA, 94105
Email: info@rockman.com
www.rockman.com

Recommended citation: Gurn, A., & Sanford-Dolly, C. (2017). Situating Big Data: Assessing Game-Based STEM Learning in Context Summative Evaluation Report. San Francisco, CA: Rockman et al.

About Rockman et al:
Rockman et al is an innovative research, evaluation, and consulting company that specializes in examining critical issues in formal and informal education. The Rockman team includes evaluators with diverse backgrounds and skill sets who help clients answer critical questions in clear, direct, and honest ways. Rockman et al has served as the lead evaluation firm for numerous projects funded by the National Science Foundation, as well as several other public and private funding agencies. Learn more at www.rockman.com.

Table of Contents

Executive Summary

Project Description

Rockman et al (REA), a San Francisco-based research and evaluation firm, conducted an external evaluation of the National Science Foundation-funded project, Situating Big Data: Assessing Game-Based STEM Learning in Context. The project was a collaboration between three research teams (two at the University of Wisconsin-Madison, and one at Arizona State University) around a shared dataset that focused on middle school youths' experiences playing *Virulent*, an online game designed to teach systems biology concepts by putting users in the role of a virus to experience and understand how viruses operate and interact within the body. Situating Big Data sought to integrate theories of situated cognition with analytic techniques derived from the big data movement.[1] Through the collection and analysis of multi-modal data generated through online and in-person interactions while youth played *Virulent*, this project explored methodological issues around the study of language and learning in the context of digital gaming.

The purpose of the external evaluation was to examine the implementation of a collaborative research design and offer an assessment of the project's methodological approaches and deliverables. To address these goals, REA used qualitative case study methods, including direct observations of the research team's data collection processes, in-depth interviews with project stakeholders, and a review of articles, presentations, and other documentation generated by the project.

Key Takeaways

This research project directly responded to a current limitation of "big data" analytics, as applied to game-based learning. Typically, learning researchers collect in-game analytics (or clickstream data) as their main metric of the effectiveness of game-based experiences. Studies of learning through online

[1] *Situated cognition* is a theoretical framework, spanning multiple disciplines, that connects social, linguistic, behavioral, and psychological dimensions of knowledge, emphasizing "the web of social and activity systems within which authentic practice takes shape (Lave, 1991, p. 84). In education, the *big data movement* refers to the application of techniques in data mining (Romero & Ventura, 2010) and learning analytics (Ferguson, 2012) so as to make use of large and complex datasets not easily managed by manual processes.

games have not merged youths' talk, social interactions, and pre/post-content assessments with analyses of clickstream "data exhaust" passively collected by the game system.

The *Situating Big Data* project explored how to curate a blended gaming environment that combined multiple qualitative and quantitative data streams, and how to effectively make use of those data streams to better understand how youth learn through games.

The research team at the University of Wisconsin-Madison designed an informal, game-based learning intervention, which served as the basis for the project's data approaches. They leveraged the digital gaming platform and group dynamics between middle school-aged children to situate youth's perceptions, gameplay, and discourse within a rich instructional context. Collected data included:

- **Pre- and Post- Surveys** to assess youths' attitudes towards science and content knowledge of cellular biology and virology,
- **Telemetry Data** to track players' clickstream pathways in the game, *Virulent*, using the Assessment Data Aggregator for Game Environments (ADAGE),
- **Discourse**, in the form of youth's verbal interactions and group discussions, and
- **Youth-created artifacts**, such as group notes and worksheets.

Although the individual data collected and the analytic strategies employed represented well-established methods, the Situating Big Data project ventured into new conceptual territory by combining these diverse approaches and theories.

Assembling and structuring this mixed dataset proved extremely challenging, though ultimately successful. One of the biggest hurdles for the project team derived from an assumption that classroom discourse could be captured in large part through an automated process. The team confronted the technological limitations of automated transcription, both in terms of data accuracy and formatting. They were forced to reallocate significant time and resources to cleaning, restructuring, and validating this data line by line. All but one project member expressed confidence in the reliability of the data. The telemetry data also required ample "human tuning" to ensure that the ADAGE system collected

analytics of interest without overburdening the processing capacity of the iPads that youth used to play the *Virulent* game.

Much of the research project was focused on developing strategies to overcome methodological problems that arose in trying to capture and connect the diverse forms of information in a single dataset. Tracking and organizing individual youth's data streams proved to be a complex and expensive process because of the quantity of data available and the amount of time required to validate the qualitative corpus. In retrospect, the research teams recognized the need to develop standard data formatting procedures at the outset of the project to more effectively manage the various data channels.

Due in part to the bottleneck caused by automated transcription issues and data restructuring needs, the research analyses across various project stakeholders started to gather momentum as the project was winding down. Utilizing the full data corpus relied heavily on the fact that team members were intellectually invested in the project and their own research goals.

Whereas strong collaborative relationships developed between two of the three teams involved in the project, one team did not play a direct role in certain key stages of the research processes and decision-making. Geographic distribution of team members, a lack of regular communication, and other external factors created constraints that limited the growth of an in-depth collaboration across all parties. Project team members recognized the ongoing work and commitment needed to build and maintain active relationships, particularly when separated by institutional and disciplinary boundaries.

Based on the teams' successes and challenges, the evaluation identified key actions that may help support future multi-disciplinary collaborations:

- *Identify specific stakeholders' needs and a driving purpose for the collaboration*
- *Recognize and plan for inherent costs and risks of collaboration*
- *Leverage existing professional relationships*
- *Develop group norms and protocols*
- *Invest in relationship building*
- *Assess progress, identify challenges, and celebrate successes*

As a result of the delays in data preparation and sharing, as well as challenges to collaboration, the project teams had not fully realized their potential in regards to empirically-grounded theory building prior to the end of the project. However, this work has demonstrated that socio-linguistic and machine learning approaches can be combined to create datasets that can be used to explore youth's learning in the context of gameplay. Further research is needed to explore the theoretical possibilities and limitations of the resulting multimodal database.

Contextualizing the Project

Over ten years ago, the Federation of American Scientists (2006) released a report maintaining that digital games presented powerful opportunities for learning. At that time, game-based learning remained largely outside of mainstream educational practice. Since then, however, the use of digital games in classroom teaching has gained traction across K-12 and out-of-school settings, as teachers, school leaders, policymakers, and parents have come to appreciate the potential educational applications and benefits of digital games. More and more practitioners recognize that digital games can provide engaging, interactive environments for formal and informal learning experiences. At the same time, advances in technology have led to increasingly complex gaming platforms that offer immersive digital worlds and collaborative multi-player gaming opportunities. The implications for learning both in and out of schools abound.

In general, game-based learning designers use visual, audio, and spatial data to construct environments that require users to make choices and take actions on problems or challenges to be overcome in order to advance to the next stage (McCall, 2012). Players are often meant to confront ill-defined or ill-structured problems that have no obvious predetermined course and may have multiple different resolutions. Players proceed individually or collaboratively in online or blended learning environments to overcome challenges and master the system. Compelling digital games motivate players to persist by presenting difficult, though achievable challenges that pique one's curiosity and provide deeply engaging experiences (Gee, 2003; Malone, 1981).

These features make digital games interesting and fun for many youth and adults, and research has demonstrated that interactive educational games can increase students' motivation to learn (Liu & Chen, 2013). However, there are many obstacles to validly and reliably assess if, what, how, and under what conditions young people are learning as a result of engaging in gameplay. This evaluation report identifies and discusses some of these barriers, and how designers and researchers sought to overcome them, in the context of analyzing youth's facilitated interactions around a science education game.

The rise of "big data" has enabled the collection and management of increasingly complex datasets from game-based learning platforms. Game designers and

researchers have developed new ways to explore users' behaviors and preferences, how youth interact with and advance through a game, and what knowledge youth gain through gameplay. Yet most data collected through educational games are "data exhaust" or telemetry data generated by clickstream behaviors, passively collected and stored as log files or cookies (Halverson & Steinkuehler, 2016). Game-generated data techniques have typically not included users' talk or discourse, a complex socio-linguistic domain underlying much of young people's learning, in and out of schools, in communities and online (Cazden, 2001; Gee, 2003, 2015).

Arguably, this failure to account for young players' discourse, relying primarily on user data collected via telemetry, obscures a great deal about the learning processes and products of gameplay. This gap in knowledge represents a collective missed opportunity for researchers to more fully understand the nature and possibilities of STEM learning in today's digitally mediated world. The *Situating Big Data* project sought to address this under-developed dimension in the research knowledge base by collecting and attempting to synthesize data from multiple sources.

Project Overview

Situating Big Data: Assessing Game-Based STEM Learning in Context is a National Science Foundation–funded project in which several university researchers partnered together to develop a multi-modal database that would capture learning through gameplay. Data of interest included youth's discourse and social interactions (in-person and online), content knowledge assessments, and in-game telemetry data. This data was collected in the context of an out-of-school program during which middle school-aged youth played the game, *Virulent*, together (see Figure 1). Youth participants' in-game and out-of-game activities were supported by adult facilitators.

Figure 1: *Virulent* Home Page

Virulent is a "serious game," meaning its primary purpose is educational rather than entertainment (Michael & Chen, 2006). Virulent was designed to teach systems biology concepts. The game puts users in the role of a virus in order to experience and understand how viruses infect cells, survive the body's immune system defenses, and replicate themselves to attack biological systems. The game can be played individually, with in-game support (an almanac) that helps users identify "characters" encountered in the game and their function to either protect the virus or defend the cell from infection. *Virulent* can be implemented in informal educational environments, such as after-school and community-based programs, or formal environments, such as middle or high school biology or health classrooms to extend or supplement the existing curriculum.

The *Situating Big Data* project team designed a multi-disciplinary research approach and developed the data plan to investigate how *Virulent* was used within an informal science learning context. With an overarching goal to promote the exchange of ideas and build knowledge across disciplines, the project's leadership envisioned a cross-institutional collaboration in which learning technologies are employed in order to:

I. Enable the creation of automated assessments that take into account the full range of pedagogical activities and discussions used to facilitate a game;

II. Better understand the interaction between player learning and the context in which that learning is transpiring; and

III. Provide rich evidence grounded in theory that can help in refining personalized learning approaches within informal and formal educational settings.

The research questions driving the project were:

- How can data streams from the contexts surrounding learning technology use be meaningfully and practically integrated into standard telemetry "big data" sets?
- What forms of analysis, from a situated cognition perspective, are enabled by the combination of in-game data, online interaction data, and in-room data?
- What new discoveries might such a "big data" situated approach to cognition enable in terms of theoretical framework refinement, methodological refinement, and the design of game-based curricula?

The project team designed an educational context to gather heterogeneous data sources, and assembled three different research teams to interpret and analyze those datasets. Two teams, located at University of Wisconsin-Madison, developed and implemented the game intervention, collected the data, and conducted both automated data mining of telemetry data through a learning analytics approach to examine young people's clickstream behavior within the digital game, and qualitative coding of player discourse to examine the verbal interactions that transpired while playing the game. A third team, located at Arizona State University, conducted linguistic analyses through a Natural Language Processing (NLP) approach to detect language patterns within the dataset and build statistical

models of linguistic features observed in the data corpus. NLP was used to analyze the hierarchical structure of the language used by youth participants over the course of the game intervention, and then to estimate the degree to which changes in participants' pre-to-post test scores could be explained by these language features.

The project sought to understand if and how these research approaches could be combined to create and make use of multi-modal data in the context of digital games.

External Evaluation Overview

Rockman et al (REA) conducted the external evaluation of the *Situating Big Data* project. The purpose of this evaluation was to explore the project team's research design, methodology, and collaborative approach to blending theories of situated cognition and big data science. For the duration of the project, REA served as a critical friend and observer, providing periodic feedback and reviewing the research team's progress towards the overall project goals (Cook & Campbell, 1979). In this way, REA offered an independent perspective on the collaborative knowledge building component of the project.

In its role as a critical friend, REA closely examined the research assumptions, methodological approaches, and intellectual products resulting from the *Situating Big Data* project using qualitative case study methods. Here, REA participated on conference calls with team members, and conducted in-depth retrospective interviews with project stakeholders from the three research teams to discuss factors that influenced how they collected, organized, and interpreted the data.

REA evaluators also observed data collection efforts at Game-A-Palooza, a camp for middle school-aged youth held on the University of Wisconsin-Madison campus during spring break. REA viewed facilitators implementing complementary educational activities around *Virulent* gameplay, as well as how the research team assessed learning around the game throughout the camp. Evaluators also conducted focus groups with Game-A-Palooza participants and interviewed facilitators about their experiences.

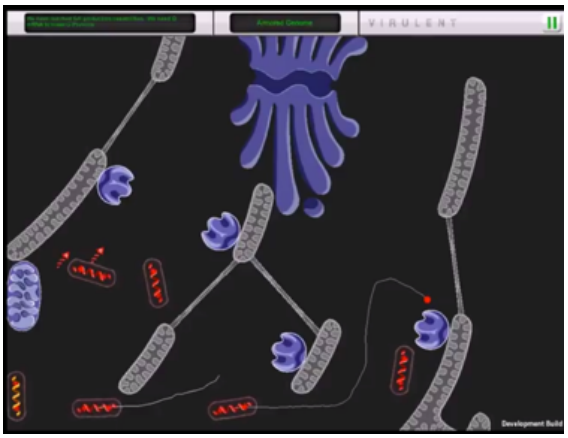
In addition, the research team provided evidence of project decisions made over time, their emerging data analysis approaches, and grant-sponsored scholarly activities and products. A literature review of collaborative presentations and publications produced by the project team was used to assess the intellectual merits and broader impacts of the research.

This summative report synthesizes the evaluation findings, responding independently to the project's research objectives, processes, and deliverables. It discusses implications for future games-based research and development projects looking to combine multiple forms of data and disciplinary perspectives.

A Serious Game that Seriously Engages

In collaboration with participating educators, the research team adapted existing curricular materials to implement *Virulent* in an inquiry-based, informal learning environment (see Figure 2). The Situating Big Data research project was not directly charged with assessing the extent to which the curricular intervention was effective. Rather, this project was focused on whether it was possible to successfully utilize a designed, game-based intervention in order to construct a dynamic dataset that could be used to explore questions about learning. However, if the game or the curricular intervention had failed to engage young people in learning, the empirical usefulness of the eventual data would have been compromised.

Figure 2: Virulent Screenshot



A real-time strategy game designed to teach systems biology, Virulent puts players in the role of the fictional, yet scientifically accurate, Raven virus. As the virus, players move through the body to infect host cells, replicate, and evade the cellular immune system. As challenges become increasingly difficult, an almanac provides just-in-time information to help identify "characters" within the game and describe their purpose to either protect the virus or defend the cell from infection.

Observations by REA evaluators at Game-A-Palooza and local, Wisconsin-based after-school programs, as well as analyses conducted internally by project team members, offered sufficient evidence that the *Virulent* game was interesting and engaging to play, and that it supported learning about systems biology (e.g., Anderson et al, 2016; Sanford & Quimby, 2016). Specifically, exposure to *Virulent* increased players' interest in and understanding of biological processes and scientific vocabulary, promoted inquiry, and scaffolded participants' critical thinking skills. The facilitated intervention extended earlier findings that young

people's engagement and learning can be enhanced through access to meaningful social interactions during gameplay (Steinkuehler, 2004; Squire & Giovanetto, 2008). Here, youth can discuss challenges they encounter during gameplay, and share strategies for overcoming those obstacles with more experienced peers and knowledgeable adults.

Youth engagement with *Virulent* derived, in part, from the fact that the instructional intervention contextualized the game in a role-playing narrative. Participants were positioned as young scientists being asked to help the CDC by investigating how the Raven virus was interacting with and destroying human cells, and coming up with strategies for how to stop it. Youth received daily updates via Skype from mock CDC scientists about the expanding Raven virus outbreak (See Figure 3).

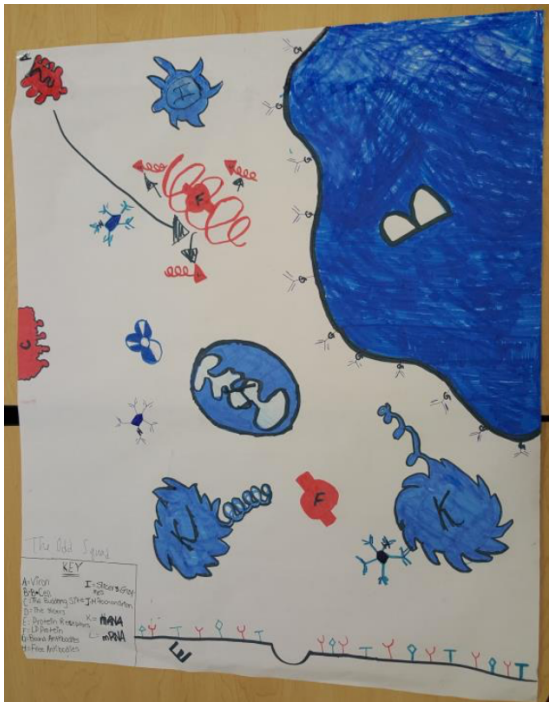
Figure 3: Skype calls with "CDC scientists" about the Raven Virus

Youth were grouped into teams of 3-4 individuals led by an adult facilitator. Team members played *Virulent* on individual devices provided by the research team, and discussed the game jointly to figure out how the virus was operating in the body. Each group worked together to tackle the challenge of stopping the Raven virus. Facilitators created daily lesson plans for the group with discussion questions focused on the game's learning objectives, and tied to the team's progress in the game and with the supplemental team challenge. Each team was asked to a) design a working model representing how the virus interacted with the host cells (see Figure 4), and b) devise a strategy to stop its spread. Each team had to explain and defend their model and proposed strategy to the whole group, with the idea that the CDC scientists would see this information later as well. These presentations took the form of explanatory videos or team-led poster

walkthroughs. Participating youth also kept notebooks, where they reflected on the game and responded to writing prompts.

The game levels and challenges within *Virulent* built on one another, becoming more difficult as players progressed through stages. On facilitator noted, *"It's like spiraling curriculum, that goes deeper as you pass. I didn't know if kids at this age level would be able to get to the very top. I was pleasantly surprised that most of them were able to."*
Analyses of pre-post youth assessments indicated that youth participants' scientific content knowledge related to virology, understanding of scientists' uses of models, and self-confidence in their ability to complete science activities increased over the week-long intervention (e.g., Anderson et al, 2016; Dalsen et al, 2016).

Figure 4: A team's working model of the Raven virus



In focus groups conducted by the external evaluators, many youth participants demonstrated knowledge of cell biology concepts and used specific vocabulary terms to describe their gameplay strategies:

Have your anti-genomes prevent the executioner from getting through.

In the final level in the third section of that level, when you have to make the virion close up the nucleus, you have to make an army.

I lost the armor, but I was able to get it back by putting the m protein. I needed to purposely lose the armor to figure it out. The slicers can't break the armored genome because it has armor on it. The protosomes eat the armor, but they can't eat the genome, so they have to work together to kill it.

Some players were motivated by the in-game challenges and showed determination in solving these problems, while other youth became frustrated when confronting particularly difficult challenges. Youth were often able to persist with the aid of in-game scaffolds like the Almanac, which provided just-in-time information to help players identify cell components and their functions. Interactions with facilitators and peers provided additional motivation to persist.

Virulent was engaging enough that many youth continued to play the game at home and during their free time. Parents shared that their children had enjoyed the supplemental curricular activities, and hoped that similar programming would be offered in subsequent years. Overall, the game-based intervention provided a stimulating learning experience for participants.

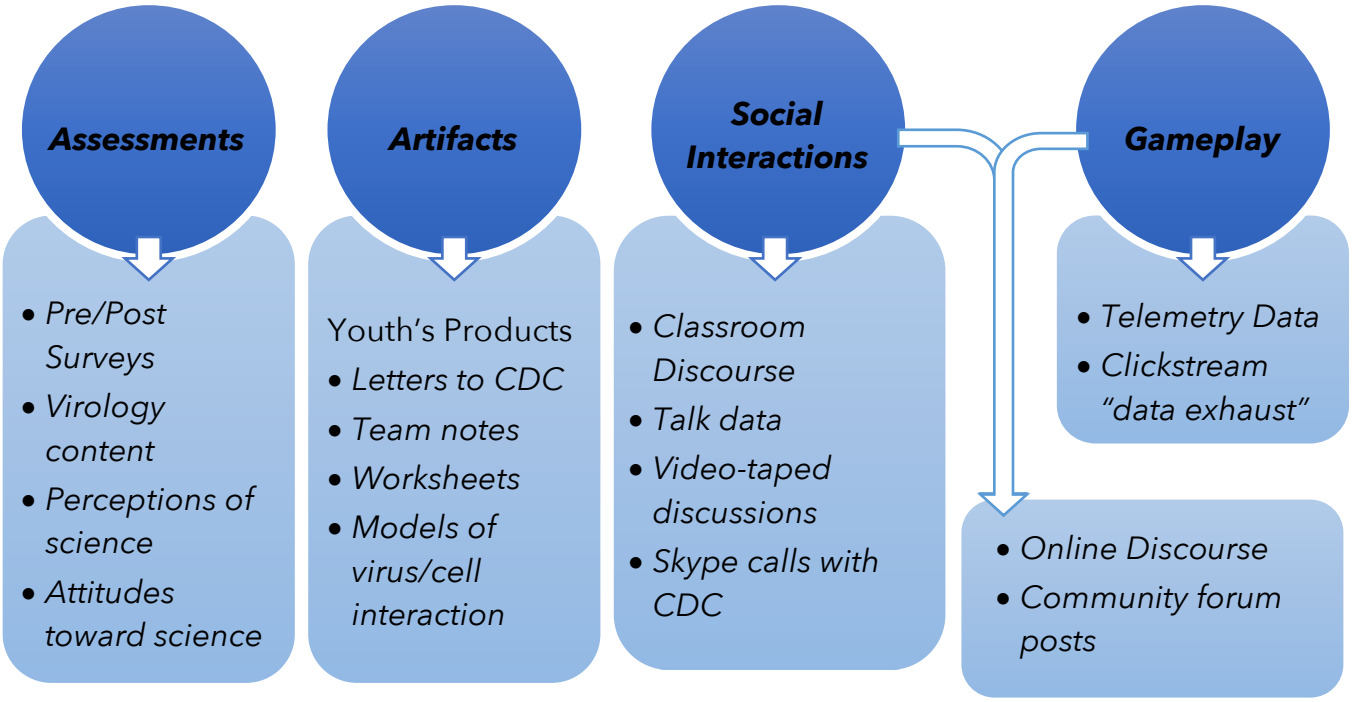
Capturing Multi-Modal Data in Context

The educational intervention designed around *Virulent* (described above) provided a basis for building the *Situating Big Data* project's data architecture. The game itself, and the group dynamics surrounding its curricular implementation, were used to document players' perceptions, behaviors, and interactions. These data collection events thus served to situate youth participants' learning and gameplay within a rich instructional context. As part of these game-based educational experiences, the research team gathered multiple forms of data from the in-person environment and the digital gaming environment (see Figure 5). These data included:

- **Pre- and Post- Youth Assessments:** Before and after the curricular intervention, youth participants took short surveys to assess their attitudes towards and content knowledge of cellular biology and virology. Survey items included multiple-choice questions, open-ended response prompts, and a diagram of cell parts to label.
- **Telemetry Data:** *Virulent* employs a backend data framework and data mining tool, called Assessment Data Aggregator for Game Environments or ADAGE (Owen & Halverson, 2013). ADAGE can be used to track players' movements within the game through telemetry or clickstream data. In *Virulent*, this "data exhaust" included: Total play time (during the intervention and overall throughout the week), game levels played, levels passed/failed, time spent on levels, challenges completed, level scores, cell resource use, Almanac resource use time spent in the Almanac, and time spent on instructions for each level.
- **Discourse:** The research team recorded youth participants' talk data (verbal interactions), while playing *Virulent* and engaging in the curricular intervention activities. Each participant wore a small audio recorder attached to his/her name badge to gather talk data. In addition, whole group activities, such as presentations and discussions about their emerging models and strategies for fighting the virus, were video recorded. Participating adult facilitators were also interviewed at multiple points throughout the week to gather their perceptions.

- **Participant Artifacts:** The research team collected and scanned documents created by youth participants each day, such as written correspondence to the CDC, group notes, and worksheets charting each team's progress towards developing their explanatory model of the virus's interactions with cells and proposed solutions for fighting the Raven virus outbreak.

Figure 5. Project Data Sources



Managing the multiple and different forms of data that were collected across each youth participant, the digital game, and the learning environment proved challenging, but not insurmountable. Collecting these data involved both the enactment of a blended educational experience and the systematic capture of socio-linguistic and online behavioral interactions. This ambitious task required coordinated efforts among multiple parties to think deeply about ways to orchestrate various forms of qualitative and quantitative data within an open-ended, game-based environment.

The *Situating Big Data* research team included veteran and emerging researchers, designers, educational practitioners, and college student interns. No single member of the team possessed a complete corpus of the professional expertise, disciplinary knowledge, and technical skills needed to effectively conduct every

aspect of the research. Several undergraduate student interns were assigned to focus exclusively on the coordination and maintenance of youth participants' data records. The project faced great difficulties in obtaining high-quality audio recordings for overlapping groups of young people moving about the space, as they played and talked about the game together. These hurdles are discussed in greater depth below.

Full analysis of the entire dataset required a range of understanding and specialized expertise across various disciplines, such as learning analytics or educational data mining, statistics, qualitative and quantitative coding, computer-based language analysis, discourse analysis, informal science learning, game-based learning, and instructional design. The research team consisted of scholars with expertise in three main schools of thought: Learning Analytics, Discourse Theory, and Natural Language Processing. The project captured diverse sources of data during gameplay in order to explore methodological questions related to the mechanisms needed to stitch these different datasets together.

Assembling researchers from multiple intellectual communities created an opportunity to learn from and integrate knowledge from a variety of disciplinary approaches. With a pragmatic eye towards the strengths and limitations of each discipline, the three research teams operated from a basic assumption that collaborative scholarship was necessary to achieve more effective study outcomes. Each of the three main disciplines are described in detail below.

Learning Analytics (LA), sometimes called Educational Data Mining (EDM), entails the measurement, collection, interpretation, and analysis of data about learners and their learning contexts (Chatti et al., 2012). As the sheer volume of educational data has grown with the development of digital technologies, research techniques from the "big data" field have been adapted for educational purposes to better understand youth learning and the environments in which that learning occurs. These approaches make use of "computerized methods to detect patterns in large collections of educational data that would otherwise be hard or impossible to analyze due to the enormous volume of data within which they exist" (Romero & Ventura, 2013, p. 12). The aim of the LA approach is to extract meaningful information from the data and utilize the results to help explain or predict learning, and/or inform curriculum development and instruction.

Digital game platforms provide researchers with the opportunity to unobtrusively collect automated data that can reveal users' in-game behaviors and interactions,

as well as to integrate out-of-game assessments of players' content knowledge. Most data collected through digital games are deemed "data exhaust" (i.e., clickstream data that is passively collected). Thus, LA techniques allow for tracking and monitoring data with little or no disruptions to natural gameplay. However, LA remains limited by the fact that digital games have tended not to capture more contextual data, such as players' talk or social-linguistic interactions, which intimately relate to the learning that transpires when young people play games (Gee, 2003). For this reason, the *Situating Big Data* project drew on theories of situated cognition, primarily through the use of discourse analysis.

Situated Cognition views learning through a socio-cultural lens, which assumes that young people's knowledge, skills, interests, and identities are formed through ongoing interactions with individuals, groups, and texts within the social and cultural contexts which they inhabit (Gee, 2015; Vygotsky, 1978). Situated cognition holds that an individual's understanding about the world is complexly shaped by social and material interactions. Discourse is one of the main ways that people engage with the world around them, and can be defined as "a socially accepted association among ways of using language, of thinking, feeling, believing, valuing, and acting that can be used to identify oneself as a member of a socially meaningful group or social network" (Gee, 1999). Cognition is thus a social activity that entails continual negotiation of meanings based in large part on linguistic interactions.

The *Situating Big Data* project adopted theories of Discourse to explore the ways that language use and social interactions relate to in-game behaviors and learning outcomes. In other words, the project explored how players' linguistic interactions might expand researchers' understanding of game-based learning beyond "data exhaust" and current LA techniques. To analyze youth participants' discourse, researchers devised a preliminary set of codes related to informal science learning and hand-coded transcripts talk-turn by talk-turn, a process that is both "laborious and time-consuming" (Creswell 2009).

Whereas traditional discourse analysis makes use of manual hand coding procedures, ***Natural Language Processing*** (NLP) is a field of artificial intelligence that utilizes machine learning to analyze the underlying linguistic features and grammatical structures of natural texts or speech language so as to explain and to predict how people understand and use language (Jurafsky & Martin, 2008). NLP is concerned with "interfacing computer representations of information with natural

languages used by humans," in order "to develop the tools for making the computers understand and manipulate the natural languages" (Khan et al, 2012).

NLP analyses result in statistical models that can detect subtle patterns in language use related to variables, such as the lexical sophistication, syntactic complexity, and semantic cohesion (Crossley, Kyle & McNamara, 2016; McNamara, Graesser, McCarthy, & Cai, 2014). In other words, NLP examines the structure of and relationships between words, sentences, and texts to determine the extent to which they hold together as meaningful units. In this study, NLP analyses were intended to help explain significant changes (or lack thereof) in young people's language use over the course of the game-based intervention, and then to predict youths' attitudes towards science, assessment results, and in-game performance based on linguistic and non-linguistic factors, such as demographics.

Ultimately, the *Situating Big Data* project sought to integrate theories and methods of situated cognition and natural language processing with the practice of educational data mining to better understand how learning takes place within digital gaming environments. Yet, collecting and analyzing data that can be combined into a multi-modal dataset is challenging, as described in the following sections.

Challenges of Data Collection, Set-up, and Processing

The project team faced numerous obstacles in structuring the multi-modal data streams, particularly with the telemetry and talk data. As one researcher put it, the team came face-to-face with the reality that "the first 90% of data analysis is cleaning, and the second 90% of data analysis is cleaning." The processes of assembling and preparing the data for analysis were critical to project success. Here, we discuss some of the challenges the team encountered with data collection and set-up.

Pre- and Post-Assessments

Before and after the curricular intervention with *Virulent*, all youth participants completed short assessments related to their knowledge about cellular biology and virology. Conducting the intervention within an informal learning context created particular constraints on the survey methods. The assessment was not positioned as a formal test, nor did it carry high stakes of any nature. Youth participants were free to skip any question, resulting in anticipated gaps in the data. Yet participants knew that the assessments were part of an important scientific research study, and appeared to approach the surveys seriously. Assessment scores were calculated as a proportion of correct responses out of the total questions answered. In addition, absences on the first or last day of the program resulted in missing data that could not be replaced.

Telemetry Data

The gameplay logs for *Virulent* were set up to record all player activity and game-controlled objects. The research team initially intended to assemble a complete account of each youth participants' clickstream behaviors within the particular gaming environment they faced. This comprehensive recording would illuminate individualized pathways and interactions within the game, with the goal of linking those in-game behaviors to potential learning outcomes.

The team chose to limit back-end recording to primarily capture player-controlled movements, only recording a game-controlled object when it died or came into contact with a player-controlled object. This choice reduced the processing load on the iPads players used, and ensured that the application could adequately respond to players' activities. In spite of this reduction, a massive amount of

telemetry data was collected. The research design team had to finesse the ADAGE system to collect the analytics of interest. As one researcher explained, these data needed to be carefully filtered:

> *Sometimes logging everything is not the thing you're looking for, and you have figure out how to filter and parse out the things that would be relevant to your interests. For example, we found pretty quickly that we didn't have automatic ways of figuring out when people had gone to the next level [of the game].*

New Python code had to be written to devise a systematic means of structuring the clickstream data, so as to provide accurate accounts of what players were doing on a near moment-by-moment basis within the game. A small handful of members on the research team performed the writing of scripts and code, and focused on devising strategies to structure, shape, and clean the gameplay data.

Ultimately, the team was successful in constructing a multi-modal dataset that situated telemetry data coming from the game with data from the facilitated instructional contexts in which *Virulent* was implemented. However, constructing the dataset in ways that satisfied the basic expectations and analytic needs of all members of the research team entailed multiple iterations and considerable investment. Organizing the telemetry data into a consistent format presented a minor hurdle in comparison to the challenge posed by the talk data.

Talk Data

The research team encountered significant problems and some setbacks when organizing the large amount of talk data that was collected. Numerous challenges arose in trying to accurately capture youth participants' discourse amidst group interactions, and then reliably linking this talk data to gameplay. Here, the project's "lamest bottleneck" became its "greatest bottleneck" (Dalsen et al, 2017).

The issue with the talk data originated with a decision made early on in the project. When planning the research proposal, the project team inadvertently overlooked a key logistical component. The team budgeted for data collection, but did not build in a sufficient budget to prepare transcriptions of the talk data for analysis. Instead, they assumed that much of this work could be done by using an automated transcription program. Indeed, figuring out ways that machine learning can facilitate applied social research was a central charge of the project. Yet in this

case, the team had to acknowledge the fact that a low- or no-cost automated solution can create more distractions than efficiency.

After compiling a batch of audio files through an automated transcription service, the team quickly realized that they would have to reconfigure the original research plan. The automated transcription produced incomplete text files with many blank spaces, inaccurate utterances, and inconsistent formatting. As one researcher explained, "Those automated services are fair at best." However, since the goal of the project was to explore language and learning across collaborating teams of learners, "Fair [did] not cut it."

One reason that the automated transcription failed was due to the fact that the small USB audio recorders that youth participants wore on their name badges did not consistently record quality sound. Since every participant wore a device and worked alongside team members, it was possible to restructure individual youth's transcripts using multiple recordings. However, this was extremely time-consuming work, much of which had to be done by hand.

The research team also faced problems associated with identifying speakers in consistent ways across the different data sources. The automated transcription services did not identify individual speakers. Transcripts these services provided noted speaker and text without discerning between different speakers (e.g., speaker 1, speaker 2, speaker 3), let alone naming individuals. Since each youth participant produced an audio file that included his or her own and multiple other participants' voices, a team of researchers had to go back and listen carefully to each recording in order to decipher *who was who* and *who said what*: "[The transcripts] were messy and they were deeply inconsistent in almost every conceivable way…There was basically a different format in every different file, and *there* were probably hundreds of files."

In addition, the research team encountered issues with inconsistent time-stamping of the transcripts. The researchers quickly recognized the importance of developing standard methods for time-stamping data that could be linked to the various data sources:

> *I don't think we realized how hard it was going to be to clean and sync the transcript data until we started seeing the transcript data.*
>
> *I think that if I restarted this project from scratch, the first thing I would do— and this sounds silly—but the first thing I would do would be [to]*

agree on a timestamp for everything we do… Every document, every pre-test, every post-test, every interview, everything to be consistent in time, and I think that would have probably shaved a year off the grant.

As a result of these challenges, the research team had to redirect significant funds into transcription. This created ripple effects in the project timeline and the nature of the work: "We had to cannibalize the grant and pour a ton of money that we didn't have into cleaning up those transcriptions by hand, not just who was speaking, but in fact what they were saying. Much of it was just left empty."

Scripts and code were developed to automate some of this data cleaning: "We spent a lot of time writing scripts to get those various transcripts into consistent, or at least reasonably consistent formats." This data structuring process required a large amount of "human tuning" that the project team had not originally allocated in terms of time, budget, and personnel.

Lastly, the project experienced a significant loss of data, including youth participant demographics, pre- and post-tests, and participant artifacts. In 2015, the Wisconsin Institute for Discovery suffered an irreparable server crash caused by severe weather that destroyed multiple research labs' data. Due to strict university IRB protocols related to the storage and transfer of research data, a large amount of data was lost. The research teams at University of Wisconsin were able to reconstruct some of those data. Other portions were excluded from analyses. To further address these issues, the research team modified some of its original expectations about the scope of the final dataset. For both financial and operational reasons, the researchers decided to limit the number of transcripts that were produced for analysis: "We ended up having to shave down a lot."

The project painfully discovered the technological limitations on sociolinguistic research in the digital age. While automated services and digital tools can reliably capture talk data for one or two speakers, there is currently no automatic or inexpensive way to distill group discourse into a text transcript without losing crucial information. Thus, sufficient time and resources must be devoted to obtain rich linguistic data in the context of group interactions.

Online Community Forum

One of the pieces of archival data collected on the back-end within the ADAGE system did not come to fruition. The project intended to have youth participants interact (i) with the game device, (ii) with others face-to-face, and (iii) with others in

an online community forum. Participants' online written posts could have offered an automated mechanism for capturing their discourse. By creating content and responding to other participants' online posts related to gameplay, the research team hoped to gain insight into users' thoughts and reflections through readily available online conversation threads that might reveal aspects of their learning.

However, participants did not use the online community in this manner. Since the project took place in informal and voluntary learning environments, the team could not require youth to participate in this space. The team initially thought participating youth would talk online with one other, sharing ideas and asking questions. Yet, the curricular activities were implemented during after-school programming and a week-long intervention, in which youth worked closely with their project teams most of the day. Many of the participants spent some time playing *Virulent* at night and during free time. However, they simply were not motivated to use the online space in a way that would provide useful data.

Although disappointing, the research team realized this was a logical feature of the *Virulent* curriculum's design. Players did not go home and log in online to talk with others about the game because their interactions during the program served this social purpose. Chatting or sharing with one another via the online game platform had little inherent draw since participants had ample opportunities to exchange ideas face-to-face with peers and program facilitators. Furthermore, establishing and maintaining user participation in online communities requires stable access to technology and regular involvement from community members and/or moderators. Thus, the community forum did not have the necessary conditions to grow.

The research team decided that the information gleaned from the online community forum was too sparse to warrant inclusion, and instead focused on the talk data. Under different circumstances, however, the result may have been very different. This method might be more effective in formal education settings, where participation can be required as part of classroom assignments or grades, or in informal settings in which the forum develops organically.

Developing Strategies for Data Analysis

The bulk of the *Situating Big Data* research project was focused on developing strategies to capture and connect diverse forms of data that could serve as valid and reliable indicators of game-based learning in context. Tracking and organizing individual participant's data streams proved to be a complex and time-intensive process due to the large amount of data collected. To conduct complete analyses, the team had to link discourse, clickstream data, youth surveys, and demographics into a coherent database.

In hindsight, the research team recognized the need to develop a data framework at the outset of the project to more effectively manage the various data channels related to individual, team, and instructional contexts. A master log of the qualitative and quantitative data was not constructed until after data collection was completed. At that stage, the team had a stronger grasp on the specific composition of the dataset. However, the timing of this data structuring, coupled with the issues experienced in obtaining quality digital transcripts, meant that researchers were cross-checking data months after collecting it. While this problem was manageable, it necessitated extra time and effort to establish a verifiable mechanism to trace the data, as well as share it across the three research teams.

Prior to initiating analysis of the data, the researchers from University of Wisconsin–Madison iteratively developed a conceptual framework based on existing research. This framework consisted of four constructs related to youth participants' learning: biology content, argumentation, modeling, and interest in science (see Tables 1-3). This *a priori* or top-down coding scheme was further refined over the course of the coding process. Here, teams of 3–4 researchers were assigned to each construct. Team meetings were utilized to raise emerging questions, negotiate construct meanings, and discuss analytic relationships. At the outset of the study, the team developed a preliminary framework of six science learning constructs (Bell & Lewenstein, 2009; Honey & Hilton, 2010) that eventually became the following four constructs:

Manual Hand Coding

Biology Content

To examine youth participants' biology content knowledge, the team developed codes for all scientific terminology related to viruses and cell biology found in the *Virulent* game and its Almanac, and in the curriculum materials (e.g., antigenome, mRNA, virion, nucleus, mitochondria, cytoskeleton, etc.). Researchers hand coded explicit uses of each term by participants, as well as implicit or vague uses, such as synonyms, descriptive statements without explicitly using terms, and mispronunciations. For each instance, they also coded for the correct use of the term.

Argumentation

To chart instances of youth participants' argumentation talk, the research team developed a coding scheme based on Berland & Reiser's (2011) research. This coding scheme accounted for individual and group instances of making claims or assertions. Argumentation codes included: Construct a Claim, Defend a Claim, Question a Claim, Evaluate a Claim, and Revise a Claim (see Table 1).

Table 1: Coding Scheme for Youth Participants' Argumentation Discourse

CODE	DESCRIPTION	DISCOURSE EXAMPLES
Construct a Claim	States a declarative fact.	"Slicers are bad." "The game is harder because the immune system is making more antibodies." "A virus is alive." "We can destroy the virus with a vaccine."
Defend a Claim	Gives an example or reason for why the claim is true	"See, look! When you hit the slicer here, the virus dies." "No, it's because there are more virions, and that's why the body is attacking them more."
Question a Claim	Asks for clarification, asks a question, asks for how someone came up with a claim	"Where do you see that?" "How can a virus think?" "How could it start at the budding site?" "What do you mean?"
Evaluate a Claim	Confirms or denies that a claim is true without offering reason as to why	"Yeah, the virus does nothing." "Look, that shield thing is protecting it." "Yeah." "No way." "That's not right."

Revise a Claim	Modifies a claim that has already been made based on new information or knowledge	"Slicers are bad for the virus, but good for the immune system." "Well, maybe a virus is alive, but it's not like... really intelligent, at least not here."

Model-Based Reasoning

To examine youth participants' development and use of models, the team developed a set of codes based on Russ et al's (2009) theoretical framework for discourse analysis of mechanistic reasoning that attends to the construction of the internal logic of the youth's model, and connections between the model and outside knowledge (see Table 2).

Table 2: Coding Scheme for Youth Participants' Model-based Reasoning

CODE	SUB-CODE	DESCRIPTION	DISCOURSE EXAMPLES
Relationships	Chaining	Reasons backwards and forwards in time	"If we remove the protein receptors, then the cell won't be able to get any proteins, and it will die." "We can't remove the mitochondria because then the cell wouldn't have any energy and would die."
	Modification	Changes, adds, refines, or removes parts of the model	"We're missing the RNA. We need to draw them here." "That's not what the genome does. We need to add these arrows." "The endoplasmic reticulum isn't an important piece. We should take it out."
Connection to Other Knowledge	Phenomena	Compares what the model says about viruses to their own experiences in the world, in order to evaluate the model.	"I have to get a flu vaccine every year, so that means vaccines aren't always effective forever. The virus must change in some way."
	Analogies	Compares virus/cell to some other system	"The mRNAs are like the running backs on the virus team. They're fast and they can avoid the slicers." "The slicers are the reinforcements, like the cavalry."

Interest

Lastly, the team devised a coding scheme to consider youth participants' interest in the game, supplemental curricular activities, and content. This last set of codes was based on Hidi & Renninger's (2008) four-phase model of interest development. Researchers coded for positive and negative value statements made by participants towards the game or activities, statements regarding the difficulty of the game or activities, and overt displays of participation, including curiosity about the content and identity statements related to participants' roles within the activities (see Table 3).

Table 3: Coding Scheme for Youths' Interest

CODE	SUB-CODE	DESCRIPTION	DISCOURSE EXAMPLES
Value Statements	Positive Negative Other	Expresses positive or negative feelings towards the game, activity, or content matter. This includes statements about the value, enjoyment, or appeal of the game or activities	**Positive:** "This music is really cool." "I like this." **Negative:** "This is boring." "I don't like this." "I don't want to do this."
Evaluating Difficulty	Easy Hard Unknown	Statements about the level of difficulty or challenge of the game, activities, or content matter.	**Easy:** "The first level was easy." **Hard:** "By the second level, I was like, 'Why is it so difficult?'"
Participation	Volunteering Persistence Withdrawing	**Volunteering:** Opting to perform a task, offer help, or share information without prompt. Includes mentoring others. **Persistence:** Showing a desire to continue a task, such as after experiencing failure. **Withdrawing:** Overt statement that child does not want to help or participate.	**Volunteering:** "I'll look it up!" "I want to do it!" **Persistence:** "I want to play just a bit longer." **Withdrawing: "**I'm not doing that."

Identity Statements		Describes characteristics or aspirations that a child has, which relate to the game, curriculum, or content matter	I am a … I want to be …
Curiosity	Exploratory Questions	Curiosity: Asking questions about the game, curriculum, or content matter.	I wonder if … What if …?

Establishing Inter-Rater Reliability

In order to ensure redundancy in the application of these four analytic constructs, approximately three months were devoted to establish inter-rater reliability. The research team randomly sampled 10-minute sections from the transcribed audio data. Two or three researchers were assigned to each of the coding schemes described above. Each researcher independently analyzed and hand coded approximately 1600 turns of talk using MAXQDA qualitative research software. Fleiss's kappa was applied to statistically measure and assess the reliability of agreement between coders. Constructs with a kappa value less than .60 were deemed to be of insufficient reliability. Team members, along with a project co-PI, met to discuss their reasoning for applying particular codes, with the aim of reaching a shared understanding over discrepancies before re-coding the discourse. This process was repeated until substantial agreement was achieved (i.e., more than .60 kappa). A graduate researcher explained how the development of the coding scheme began before the data collection events and proceeded throughout analysis:

> *Prior to analysis, our team would meet every week, in order to make sure that everyone was on the same page, to give updates, so that the coding schemes got feedback, and if there were any questions or concerns, to be able to address those as we were leading up to data collection. After data collection, we continued that process… The weekly meetings with the team became bi-weekly just because we were coding, and then we had our [whole group] meetings, where people could also ask questions or talk about updates.*

Automated Coding

Computational Analysis of Telemetry Data

The telemetry data from ADAGE were also analyzed using two of the four learning constructs described above (biology content and interest). However, instead of language examples, these codes were operationalized in terms of identifiable behaviors in players' clickstream data. Table 4 shows the coding scheme, as applied to the telemetry variables tracked and monitored in *Virulent*.

Table 4: Coding Scheme with Telemetry Data

CODE	TELEMETRY VARIABLE
Biology Content	Time Spent in Almanac; Almanac Entries Referenced Play Time Totals Levels Played; Level Success (pass/fail, time taken) as indicator of exposure to content Scores per Level
Argumentation	Not Applicable
Modeling	Not Applicable
Interest	Play Time Totals (amount played in-program and out-of-program) Challenge Completions (judged by coin rewards' criteria) Number of attempts or replays before succeeding at a level

Computational Analysis of Natural Language Data

The transcripts of youth participants' discourse were separated by individual youth and run through three freely-available natural language processing (NLP) tools so as to isolate linguistic information related to text cohesion, lexical sophistication, and sentiment analysis (see Table 5).

Table 5: Natural language processing tools

NLP TOOL	DESCRIPTION
Automatic Analysis of Lexical Sophistication (TAALES)	TAALES is a NLP tool used for batch processing of text files to examine over 150 indicators of lexical sophistication, for example, measuring word frequency and frequency of academic words and phrases, and other indices (Kyle & Crossley, 2015).

Tool for the Automatic Analysis of Text Cohesion (TAACO)	TAACO measures over 150 linguistic features related to text cohesion and lexical variation, such as, type/token ratio indices that measure the total number of different words (types) by the total number of words overall (tokens), sentence overlap indices that assess local cohesion, paragraph overlap indices that asses global text cohesion, and use of connective devices [e.g., but, however, in spite of] (Crossley, Kyle, & McNamara, 2016).
SEntiment ANalysis and Cognition Engine (SEANCE)	SEANCE is a tool to conduct analysis of semantic information related to feelings and opinions. Based on a variety of preexisting databases, SEANCE assesses positive and negative sentiment, cognition, and social order (Crossley, Kyle, & McNamara, 2017).

Connecting Analyses Across Data Streams

Linking the telemetry and talk data has shown promising points of inquiry. For instance, the research team has identified a correlation between science content knowledge and the number of times a player failed at a particularly difficult level of the game. A graduate researcher explained:

> *There was one level that was specifically designed to be difficult. This was part of their first boss level, where prior to this, they're being taught how to play the game. And this level is the, 'Now you know how, let's see you do it.' …The more that they failed at this level before completing it, the more they learned overall in the event.*

Analyses for the *Situating Big Data* project were currently on-going. At present, the research teams at the University of Wisconsin–Madison are digging into the meaning and utility of failure in critical thinking skills and creativity (Anderson, Kumar, Dalsen, Berland & Steinkuehler, in revision). They are considering how players were changed as a result of struggling through and then succeeding at a difficult level of *Virulent*, as well as the mediating role of where this level was situated in the game and of the curricular intervention. More specifically, the group is "trying to figure out if there is something special about putting a designed difficult level in an educational game that can help get that embedded material across."

As implemented, the *Virulent* intervention walked youth participants through the early levels of gameplay to help them become acquainted with the game and related biology content. Completing the earlier levels of the game was not simple, and players did experience some failure before they reached the first 'boss level.'

However, this difficult stage showed youth that the game was going be challenging, suggesting to them that, "You're gonna have to try. You're gonna have to pay attention and you're gonna have to figure out how to do this." The adult facilitators were trained to not give the youth participants answers, but rather to help guide their thinking and problem-solving. The research team is currently exploring the extent to which participants' persistence through failure influenced their interest and learning outcomes, and identifying potential underlying explanations (Anderson et al, in revision).

From a Natural Language Processing (NLP) approach, the researchers at Arizona State University focused on language features in the transcripts of youth participants' discourse with the aim of linking those latent language features with the project's learning outcomes. One researcher explained:

> *Our theoretical assumption is that not only do the words that you produce provide information about [learning], but also the features of those words and features of the sentences and the overlap between ideas. The underlying notion is to take various levels of language related to an outcome, and then look at the degree to which those features of language, not the word that you say, but the features of what you're saying, predict outcomes…The notion is that the language we produce is a proxy for these various theoretical constructs.*

The NLP analysis looked to connect youth participants' use of language features (e.g., text cohesion, lexical sophistication, and sentiment) to virology content, as evidenced in their pre- and post- assessments. Members of the Arizona State University research team felt the project was moderately successful in this respect. In one study, researchers observed a statistically significant correlation and medium effect size between the learning outcomes and the relative sophistication of youth participants' language features, after controlling for non-linguistic factors, including gender, age, and ethnic group, as well as prior experience with technology, gaming preferences, interest in science, and favorite school subject (Crossley et al, under review). Linear mixed effect models showed that both linguistic factors (R^2 = .525) and non-linguistic factors (R^2 = .482) were predictive of participants' science test scores. Linguistic variables were better predictors of science scores than non-linguistic variables. A combined model, which accounted for both non-linguistic and linguistic factors, explained approximately 61% of the variance in the science scores (R^2 = .609). Students with stronger linguistic skills tended to exhibit higher science gains. In addition, demographic variables such as gender and ethnicity were not predictive of science tests scores, suggesting that

females, males, and all ethnic groups performed similarly. One researcher explained that the implications of this NLP analysis can be interpreted in a number of ways:

> *You can basically interpret the findings and say, 'Those people who have greater languages skills, those language skills themselves allow them to collaborate better, to understand difficult domains, to acquire new knowledge, and as a result they're more successful at something like learning about virology.*

In other words, youth with greater command of the language are better able to grasp the science content matter and subsequently perform better on assessment activities.

Alternately, the team has explored the theory that there is a cognitive domain that underlies one's ability to engage in analyses of complex problems or phenomena, in general, which in turn influences how a person performs in science: "At the same time, that ability to analyze things also allows you to be more proficient or less proficient at language." In other words, young people's creative and critical thinking skills may impact their ability to engage in linguistic activities, such as reading, writing, discussion, and debate, which are embedded within science learning contexts.

At this point, the research teams can say with confidence that there is a link between the sophistication of youth participants' language use and their science knowledge. The project can demonstrate that a relationship exists, although the specific nature of that relationship is not yet known. To explore the connections between these variables, more research is needed to parse out how language use, non-linguistic factors, and the learning of scientific ideas intersect.

Next Steps for Analysis

After having to navigate the multiple theoretical and operational obstacles discussed above, the team's research analyses really picked up momentum as the project began to wind down. As the PI's moved toward synthesis across the full data corpus, they have relied heavily on the fact that a core group of team members are intellectually invested in the project and potential future research. In the final year of the project, funding ran out to pay graduate students for their time. Graduate student researchers have continued their work on the project largely

because it was part of a paper that they wanted to co-author or formed the basis of a dissertation. As two researchers put it:

> *Imagine our budget is done, we have no more students on this, and what we're doing now is trying to finish up analyses…on good graces and on the fact that we're intellectually interested, not because we have students to pay to do it… There is no more money. We are all working on analysis and papers in order to produce the work we promised.*
>
> *Just because of it being interesting to students, people are invested in the project, people are still writing.*

In spite of the multiple challenges that slowed and modified research plans, there have been notable project accomplishments in terms of research findings and theory building. The *Situating Big Data* project helped generate numerous peer-reviewed conference presentations, papers for edited volumes, journal articles in preparation or revision, a dissertation proposal, and supported the development of publicly-available language research tools (See Appendix: Project-Supported Scholarship).

However, the obstacles faced early on limited some of the project's impact on the field. However, these limitations may be a matter of shifting timelines, rather than methodological expectations. One researcher noted that the project is still working towards meeting its objectives: "Overall, I think that we haven't fully reached our potential yet for this project. I really do have high hopes for the things we're going to be able to produce in the near future."

The Promise and Pitfalls of Multi-Disciplinary Research Teams

Critical reflections by project team members illuminated the complicated and delicate nature of conducting collaborative research that spans disciplines, methods, and geographic locations. The team navigated multiple methodological issues that revealed both promising practices and substantial challenges to effectively and efficiently conduct collaborative research in this domain.

Over the past decade, engineering, behavioral, and social scientists have increasingly been encouraged to engage in collaborative, multi-disciplinary work. More and more, funders expect researchers and practitioners to reach beyond their own discipline and work with scholars from varying fields. While working with multiple researchers from different disciplines can enhance the diversity of thought and expand collective expertise, the decision to engage in collaborations also comes with time and resource costs to coordinate researchers, calibrate theoretical assumptions, agree upon methodological approaches, and develop joint products.

Given the rise in support and expectations for multi-disciplinary teams, it is important to better understand the inter-related factors that may enhance or hamper researchers' engagement and participation in collaborative efforts. The National Research Council (2015) promotes this view, arguing that "research is needed to enhance our basic understanding of team science processes as the foundation for developing new interventions" (p. 12).

In the *Situating Big Data* project, two out of the three research teams collaborated actively, engaging in cooperative efforts towards a shared goal. Project members in these two groups found cross-team interactions to be intellectually stimulating and productive. The third team, located at a separate university in another state, generally expressed feelings of detachment, and felt uninvolved in the research processes, decision-making, and collaborative analyses. All project team members interviewed as part of this evaluation agreed that the full-blown collaboration did not materialize as intended across the three teams, although there was some disagreement about the reasons for this breakdown. At the outset of the project, no one anticipated experiencing major obstacles to communication and collaboration. However, numerous takeaways emerged from reflection.

Aspects of a Successful Collaboration

For the two teams located on the same campus, the collaboration was led and facilitated by two professors (co-PI's on the project) who worked closely to recruit and assemble graduate student researchers with appropriate experience and interest to successfully participate in the project. Initially, graduate students received stipends to meet regularly as a group with the PI's. When a team member ran into a problem with how data files were organized or how particular codes were being conceptualized, these questions were addressed within the structure of a weekly meeting.

Meetings were used to iron out conceptual differences between how the two research teams viewed data and the analysis of data. These differences were not trivial and required time to ask probing questions, to critically reflect on disciplinary assumptions, and to develop shared research perspectives. For example, educational data scientists and situated cognitive theorists may draw on related, but diverging sources of knowledge and incumbent terminology to explain the same concept. In practice, the teams made different assumptions about what youth participants' interest would look like across the different datasets. Ongoing dialogue was necessary to calibrate these assumptions, compare data, and come to common understandings of key concepts.

For instance, the team knew that some youth participants were identified as having special learning needs. In working to understand players' individualized pathways through the game, the groups discussed if and how it would be possible to differentiate youth who may have a very different set of learning needs. One researcher explained:

> *We had conversations about what [learning] looks like in the game files in a quantitative sense. What does it look like in the talk files in a qualitative sense? And how would you find moments of connecting them?*

Productive discussions resulted from the fact that the two teams were charged with having to organize and manage one dataset and explain it to the other team. One team had greater command of the telemetry data. One team focused on youth participants' talk data. When either team had questions about the other team's dataset, they had to ask for help to interpret the information, which led to analytic discussions:

Our two teams were working kind of cheek to jowl. Every time we ran into a problem with how the data files are organized, we were working it out on a weekly basis…differences between how we view data, how we view analysis of data.

When we were coming up with the constructs [used in the coding scheme], there were so many interesting discussions. We dug into the related literature and we talked a lot about connections we saw to the project. Like, what kinds of evidence we wanted to look for. It helped that many of us came from different fields.

Challenging One Another's Assumptions

While researchers within the same field often disagree in their understandings and perceptions of the particular nature of a research problem, multi-disciplinary teams face the added burden of approaching research with different underlying languages, different theoretical assumptions, and different methodological tools. The inter-disciplinary conversations described above challenged each team's research presumptions. For example, in analyzing youths' discourse, the teams developed skepticism about counting instances of talk data and extrapolating meaning. In a quantitative data set, the greater number of instances of something often entails greater likelihood of finding a statistically significant relationship or observing a larger effect size. On the other hand, in the case of talk data, "A single line of someone saying, 'I hate you and I never want to be here again,' is one turn of talk, but it can change everything."

One graduate researcher recalled several exchanges in which working with researchers from other disciplines forced the group to be more explicit about meanings. At one point, two engineering students participated in the project who looked at the research from a different, "more procedural" perspective. When graduate students in education referred to teaching and learning concepts or a particular theoretical framework, the engineering students would ask for clarification or pose questions. The ensuing conversations pushed the team members to explain operating definitions and re-examine their own assumptions about the game and the learning context in which it was situated.

While the two teams at University of Wisconsin–Madison used differing methodologies and theoretical assumptions, they were able to work through these differences because of their shared curiosity in the research problem and ample time allotted to reflect on the data together. Their collaborative efforts could be

seen as a successful form of *interdisciplinary research* that integrates "information, data, techniques, tools, perspectives, concepts, and/or theories from two or more disciplines . . . to advance fundamental understanding or to solve problems" (National Academy of Sciences, National Academy of Engineering, and Institute of Medicine, 2005, p. 26).

Obstacles to Collaboration

Trying to stitch together and analyze very heterogeneous data was a complex undertaking. Doing so across long distances among researchers without an established history of working together confounded matters further. The teams located on one campus had the benefit of face-to-face exchanges to jointly address issues when they arose: "We were co-located, so we had this luxury of being able to sit in-person with each other and figure out why something was not working." In contrast, other communication channels, such as phone calls and email, were needed for remote collaboration with the team located at another institution in another state. These communication channels were not well-maintained over the course of the project.

The success of the first two teams' collaborative efforts to cross disciplinary boundaries derived in part from ongoing discussions that facilitated the development of a common vocabulary and a shared conceptual framework to think about games, language, and learning. Ironically, the essential theory driving the project, that much of learning transpires through social interactions, was not actualized across all three teams. As one researcher put it:

> *Our intellectual work all comes down to inter-personal relationships. You would think I'd know that because, as a socio-cultural researcher, I know that the social interaction drives learning. Yet, I really did not see that coming.*

Another issue that hindered multi-disciplinary collaboration was that the third research team was not intimately involved from the start of the project. They did not engage closely in many of the initial conversations about the research process, such as refining the overall research plan, setting up collaborative processes, conducting data collection, building the dataset, or conceptualizing key constructs. From a logistical standpoint, they were positioned in the original project proposal more as expert research consultants and subcontractors, rather than as partners.

The third team did not play a direct role in developing the research protocols and did not initially provide input on what essential data and data formats they needed to conduct their respective analyses. This caused problems later in the project that might have been avoided, if all stakeholders were involved in the initial design process. One researcher joked:

> *In the future, I'm going to make everybody sign in blood consistent data formats. On day one, we are going to all have to become blood siblings around data formats and data structures. That's my biggest takeaway.*

The lack of consistent communication between parties or a clear structure to regularly share ideas and issues that arose across all three teams created obstacles to collaboration. For these reasons, research team members from all three groups agreed that the *Situating Big Data* project did not achieve all the innovative discoveries that they had hoped. However, since analysis and writing up of the methodological issues and research findings are still very much underway, there is a lot more that can be done with the research products.

Features of Successful Multi-Disciplinary Collaboration

Based on the evaluation findings above, several key takeaways have been identified that may help support future multi-disciplinary collaborations:

- ***Identify specific stakeholders' needs and a driving purpose for collaboration***

Multi-team research collaborations may be moderated by university-, industry-, or funder-related expectations. As such, researchers must spend sufficient time at the beginning of the project to convene potential partners and develop shared understandings of the overall goals of the project, potential intellectual contributions, and intended deliverables.

- ***Recognize and plan for inherent costs and risks of collaboration***

Initiating and maintaining a multi-disciplinary collaboration entails significant investments of time and resources to develop effective and efficient relationships among scholars trained in different ways of thinking and researching. Furthermore, while many organizations promote interdisciplinary research, they may offer limited resources to strategically support these collaborative relationships. Fair and pragmatic acknowledgement of the anticipated costs and benefits, as well as identification of potential threats to success may help to avoid later pitfalls.

- ***Leverage existing professional relationships***

When assembling a team, it is vital to have the appropriate mix of researchers with expertise and experience that fits the nature of the project. This may require seeking out and forging new relationships. However, utilizing existing relationships with individuals who have a good track record of working together may reduce the start-up time associated with developing rapport.

- ***Develop group norms and protocols***

It is important for collaborators to mutually establish a set of shared expectations for communication and collaboration. Individuals may hold different notions of effective behaviors, or expect interactions to play out in particular ways. Directly address the partnership terms at the outset, for example, discussing how team members will work together, key roles and responsibilities, mechanisms and expectations for communication, how decisions will be made, how eventual conflicts will be handled, and how authorship will be determined. Although issues or problems usually require individual attention when they arise, having these conversations up front will help to establish transparent processes, and ensure that all parties understand what is expected of them.

- ***Invest in relationship building***

Collaborations cannot succeed without productive working relationships. In the early stages of a collaboration, particularly one in which participants do not have history working together, face-to-face interactions may be necessary to build trust and shared understanding. If face-to-face interactions are not feasible, frequent telephone or video conferencing check-ins throughout all stages of the project can help team members to feel included and involved in project discussions and decision-making. In other words, invest ample amounts of time into getting to know one another professionally.

- ***Assess progress, identify challenges, and celebrate successes***

Research teams need transparent means of defining their objectives, as well as a way to determine when those goals are met. Anticipate setbacks and challenges, and negotiate strategies to overcome them. Create opportunities for early "wins" to galvanize support and buy-in across team members. Acknowledge individual and collective efforts and openly communicate successes, as well as areas in need of improvement. Such reflective practices can help solidify the group as a functional team, and motivate individuals to engage in additional collaborations in the future.

Conclusion

The *Situating Big Data* project began with the premise that well-structured digital games can support players' learning. However, the prevailing research tools and practices drawn from the "big data" movement do not account for group discourse that naturally occurs as young people play games. Failing to capture youth's talk and social interactions alongside gameplay creates a narrow view of learning processes. Therefore, the growing anticipation and enthusiasm regarding the intellectual or educational merits of game-based learning analytics may be inflated. This project explored ways to avoid losing important contextual data related to youth learning in games, which is not represented through clickstream data exhaust.

The research team developed an educational intervention around the game, *Virulent*, within an informal learning setting. The project captured multiple sources of data from gameplay, including log data, clickstream data, talk data, and youth-produced documents, in order to explore what it takes to stitch together different datasets. In doing so, the team sought to marry theories of situated cognition with educational data mining.

Ultimately, the project was successful in constructing a multi-modal dataset that situated telemetry data coming from the game with data from the facilitated programmatic contexts in which the technology was implemented. However, there were numerous methodological challenges that arose in assembling and structuring these data, which may serve as a cautionary tale for other researchers.

First, the project team confronted the reality that current accessible technology does not effectively offer an automated means to produce transcriptions of multi-player language interactions without losing vital data related to learning in context. While low- or no-cost automated programs can adequately distill speech from one or two individuals, educational environments—especially interactive ones—require more refined interpretation of group conversations. In the future, similar projects should allocate considerable resources to transcribing verbal interactions. Alternately, it may possible to use telemetry data, field note observations, and/or youth or educator interviews to help identify areas of interest within gameplay. Here, specific portions of audio recordings can be targeted to explore a study's object of interest.

Second, the project encountered major hurdles stemming from the fact that the research teams did not develop and agreed upon standard data formats and structures at the outset of the project. Significant time and resources had to be poured into establishing consistent means for identifying speakers in audio files and time-stamping the various data streams. Prior to data collection, other researchers are advised to devise a master data log to track and organize all data sources.

Third, the project team found that not all planned data sources readily fit the learning context. For example, players did not have an inherent reason to make use of an online community forum, which would have offered an automated mechanism for analyzing written discourse. Since the intervention provided ample social interactions around gameplay, there was no pressing need for youth to interact around the game online. While this data-gathering method did not flourish in this specific research context, other scenarios may elicit greater effectiveness, such as formal education settings, where participation is required or part of one's grade.

Lastly, this project demonstrated how the capacity to enact inter-disciplinary scholarship rests largely on the ability of those involved to build and maintain relationships. These kinds of collaborations require ongoing joint activities and trust-building. This did not happen across all parties in the project due to geographic dispersion, a lack of regular communication between project partners, and other external factors. Similar future work would benefit from the joint design of a system that establishes procedures for team communications, setting group norms, reviewing inter-team processes, and consensus decision-making.

Multi-modal datasets offer enticing avenues to assess the role and impact of digital games on learning. To realize this possibility, new methodological approaches and tools are needed that can accurately capture the complexity of learning that transpires within game-based environments. This project showed that socio-linguistic and machine learning approaches can be combined to create actionable data that explores youth learning in the context of gameplay. At this time, computational methods cannot replace manual methods of interpreting and analyzing highly contextualized social experiences. Furthermore, computational techniques continue to require significant human tuning to ensure proper fit between designed analytics and the research context. A mixed approach offers promising solutions for examining multiple data channels related to learning.

However, further research is needed to fully explore the theoretical potential and limitations of the multi-modal database developed through the *Situating Big Data* project.

Citations

Alexander, R. J. (2004). Towards Dialogic Teaching: Rethinking Classroom talk. Cambridge: Dialogos.

Anderson, C., Dalsen, J., Stenerson, M., Robinson, J., Binzak, J., Wielgus, L., Ebert, S., Azari, D., Bloker, L., Scaico, P., Bohanen, R., Squire, K., & Steinkuehler, C. (2015). Game-a-Palooza. 11th Annual Games + Learning + Society Conference, Madison, WI. GLS 11 Conference Proceedings. MIT Press: Cambridge.

Anderson, C. G., Binzak, J. V., Dalsen, J., Saucerman, J., Jordan-Douglass, A., Kumar, V., Turker, A., Scaico, P., Scaico, A., Berland, M., Squire, K., & Steinkuehler, C. (2016). Situating Deep Multimodal Data on Game-Based STEM Learning. Proceedings from ICLS 2016: 12th International Conference of the Learning Sciences.

Bell, P. & Lewenstein, B. (2009). Learning science in informal environments: People, places, and pursuits. Washington DC: National Academies Press.

Berland, L. K., & Reiser, B. J. (2011). Classroom communities' adaptations of the practice of scientific argumentation. Science Education, 95(2), 191-216.

Cazden, C.B. (2001). Classroom Discourse: The Language of Teaching and Learning, Portsmouth, (2nd ed.), NH: Heinemann.

Creswell, John W. 2009. Research design: Qualitative, quantitative, and mixed methods approaches. Thousand Oaks, CA: SAGE Publications.

Crossley, S., Allen, L. K., Snow, E. L., & McNamara, D. S. (2015). Pssst. . . Textual features. . . There is more to automatic essay scoring than just you! Paper presented at the Fifth International Conference on Learning Analytics and Knowledge, Poughkeepsie, NY.

Crossley, S. A., Kyle, K., & McNamara, D. S. (2017). Sentiment analysis and social cognition engine (SEANCE): An automatic tool for sentiment, social cognition, and social order analysis. Behavior Research Methods 49(3), 803-821.

Federation of American Scientists. (2006). Report: Summit on Educational Games: Harnessing the Power of Video Games for Learning. Washington, D.C.

Ferguson, R. 2012. The State of Learning Analytics: A Review and Future Challenges. Technical Report KMI-12-01, Knowledge Media Institute, The Open University, UK.

Gee, J. P. (1999). An introduction to discourse analysis: Theory and method. London: Routledge.

Gee, J.P. (2015). Social linguistics and literacies: Ideology in discourses. New York: Routledge.

Gee, J.P. (2007). What video games have to teach us about learning and literacy. New York: Palgrave MacMillan

Halverson, R., & Steinkuehler, C. (2016). Games and Learning. *The SAGE Handbook of E-learning Research*, 375.

Hidi, S., & Renninger, K. A. (2006). The four-phase model of interest development. Educational Psychologist, 41(2), 111-127.

Honey, M. A., & Hilton, M. (Eds.). (2010). Learning Science through Computer Games and Simulations. National Research Council. Washington, DC: National Academy Press.

Khan, M., Dar, M., & Quadri, S. (2012). Towards Understanding Theoretical Developments in Natural Language Processing. International Journal of Computer Applications, 38(2), 1-5.

Lave, J. (1991). Situated learning in communities of practice. In L. B. Resnick, J. M. Levine, & S. D. Teasley (Eds). Perspectives on socially shared cognition (pp. 63-82). Washington, DC: American Psychological Association.

Liu, E. & Chen, P. (2013). The Effect of Game-Based Learning on Students' Learning Performance in Science Learning – A Case of "Conveyance Go. Procedia, 103, 1044 – 1051.

Malone, T. (1981). What makes computer games fun? Byte, 6(12), 258–277.

McNamara, D. S., Graesser, A. C., McCarthy, P. M., & Cai, Z. (2014). Automated evaluation of text and discourse with CohMetrix. New York, NY: Cambridge University Press.

Michael, D. & Chen, S. (2006). Serious games: Games that educate, train, and inform. Boston, MA.: Thomson Course Technology.

National Research Council. (2015). Enhancing the Effectiveness of Team Science. Committee on the Science of Team Science, N.J. Cooke and M.L. Hilton, Editors. Washington, DC: The National Academies Press.

Romero, C. & Ventura, S. (2010). Educational Data Mining: A Review of the State of the Art. IEE Transactions on Systems, Man, and Cybernetics-Part C: Applications and Reviews, 40(6): 601–625.

Russ, R. S., Coffey, J. E., Hammer, D., & Hutchison, P. (2009). Making classroom assessment more accountable to scientific reasoning: A case for attending to mechanistic thinking. Science Education, 93(5), 875-891.

Sanford, C, & Quimby, C. (2016). *CyberSTEM Summative Evaluation Report*. San Francisco, CA: Rockman et al.

Appendix: Project-Supported Scholarship

Conference papers/posters/presentations (Juried):

Anderson, C., G., Binzak, J., V., Dalsen, J., Jordan-Douglass, A., Kumar, V., Turker, A., Saucerman, J., Berland, M., & Steinkuehler, C. (2016). Connecting Gameplay, Discourse, and Assessment in a Learning Game Camp. *Proceedings of the 12th Games + Learning + Society Conference*. Madison, WI. USA.

Anderson, C. G., Binzak, J. V., Dalsen, J., Saucerman, J., Jordan-Douglass, A., Kumar, V., Turker, A., Scaico, P., Scaico, A., Berland, M., Squire, K., & Steinkuehler, C. (2016). Situating Deep Multimodal Data on Game-Based STEM Learning. *Proceedings of the 12th International Conference of the Learning Sciences* (ICLS-16). Singapore.

Anderson, C.G., Dalsen, J., Stenerson, M., Robinson, J., Binzak, J., Wielgus, L., Ebert, S., Azari, D., Meschke, V., Bloker, L., Bohanan, R., Scacio, P., Squire, K. & Steinkuehler, C. (2015). Game-a-Palooza: Games, Fun, Learning. *Proceedings of the 11th Games + Learning + Society Conference*. Madison, WI. USA.

Binzak, J., V., Anderson, C., G., Kumar, V., Jordan-Douglass, A., Berland, M. (2016). Comparing Gameplay Across Formal and Informal Contexts. *Proceedings of DiGRA-FDG 2016*. Dundee, Scotland.

Crossley, S., McNamara, D., Dalsen, J., Anderson, C., Berland, M., & Steinkuehler, C. (under review). Linking natural language to science. In Conati, C., Heffernan, N., Mitrovic, A., & Verdejo, M.F. (eds.). *Proceedings of the Artificial Intelligence in Education (AIED) Conference*. Heidelberg, Germany: Springer.

Stenerson, M., Turker, A. (2016). Visualizing Game Data: Collaborative Dashboard Design for Researchers and Teachers by Researchers and Teachers. *Proceedings of the 12th International Conference on Games + Learning + Society Conference.* Madison, WI. USA.

Anderson, C. G., Berland, M., Binzak, J. V., Dalsen, J., Jordan-Douglass, A., Kumar, V., Saucerman, J., Turker, A. & Steinkuehler, C. (2016). Connecting Gameplay, Discourse, and Assessment in a Learning Game Camp. *Proceedings of the 12th Games + Learning + Society Conference*. Madison, WI. USA.

Steinkuehler, C., Berland, M., Squire, K., Anderson, C. G., Binzak, J. V., Wielgus, L., Azari, D., Dalsen, J. & Scaico, P. (2015). Situating Big Data Across Heterogeneous Data Sets of Game Data Exhaust, Class Assessment Measures, and Student Talk. *Proceedings of the 11th Games + Learning + Society Conference*. Madison, WI. USA.

Conference papers (Not Juried):

Anderson, C. G, Dalsen, J., Kumar, V., Berland, M., Steinkuehler, C. (2017). Failing Up - The Role of Difficulty and Failure in an Educational Video Game. Presented at the 2017 Digital Media and Learning Conference. Irvine, CA. USA.

Journal articles:

Anderson, C. G., Kumar, V., Dalsen, J., Berland, M. & Steinkuehler, C. (in revision). How Failure Promotes Learning Through Discourse in Game Environments. *Thinking Skills and Creativity*.

Dalsen, J., Turker, A., Berland, M., & Steinkuehler, C. (in progress). Constructing and tracing scientific argumentation through gameplay. To be submitted to *Science Education*.

Workshops:

Turker. A., & Dalsen. J. (2017). Challenges to Multimodal Data Set Collection in Games-based Learning Environment. *Current and Future Multimodal Learning Analytics Data Challenges*.

Book Chapter:

Dalsen, J., Anderson, C. G., Squire, K. & Steinkuehler, C. (in press) Situating Big Data In M. Young & S. Slota (Eds.), *Exploding The Castle: Rethinking How Video Games & Game Mechanics Can Shape The Future Of Education* (pp. 216-242). Charlotte, North Carolina. USA.

Dissertation work:

Dalsen, J. (in progress, expected 2017) Constructing the language of science: Argumentation, gameplay and the diverse learner. (PhD dissertation). University of Wisconsin–Madison.

Work on Tools to conduct analyses:

Kyle, K. & Crossley, S. A. (2015). Automatically assessing lexical sophistication: Indices, tools, findings, and application. *TESOL Quarterly 49*(4), pp. 757-786.

Crossley, S. A., Kyle, K., & McNamara, D. S. (2017). Sentiment analysis and social cognition engine (SEANCE): An automatic tool for sentiment, social cognition, and social order analysis. *Behavior Research Methods 49*(3), pp. 803-821.

Crossley, S. A., Kyle, K., & McNamara, D. S. (2016). The development and use of cohesive devices in L2 writing and their relations to judgments of essay quality. *The Journal of Second Language Writing, 32, 1-16*.

Conference papers (In Progress):

Dalsen, J., Turker, A. G., Kumar, V., A., Berland, M., Squire, K., & Steinkuehler, C. (in progress). *Proceedings of the 13th International Conference of the Learning Sciences* (ICLS-18). London.