

PLAYFUL TESTING



DESIGNING A FORMATIVE ASSESSMENT GAME FOR DATA SCIENCE

NATHAN HOLBERT • DAISY RUTSTEIN • MATTHEW BERLAND
BETSY DISALVO • JEREMY ROSCELLE • VISHESH KUMAR
SATABDI BASU • REINA FUJII • BETH PINZUR

Playful Testing

PLAYFUL TESTING

*Designing a Formative Assessment Game for Data
Science*

NATHAN HOLBERT, DAISY RUTSTEIN, MATTHEW
BERLAND, BETSY DISALVO, JEREMY ROSCHELLE,
VISHESH KUMAR, SATABDI BASU, REINA FUJII, &
BETH PINZUR

Carnegie Mellon University: ETC Press

Pittsburgh, PA



Playful Testing by Carnegie Mellon University: ETC Press is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, except where otherwise noted.

Copyright © by ETC Press 2022 press.etc.cmu.edu

ISBN: 978-1-387-50483-1 (Print)

ISBN: 978-1-387-47594-0 (ePUB)

The text of this work is licensed under a Creative Commons Attribution-NonCommercial-NonDerivative 2.5 License (creativecommons.org/licenses/by-nc-nd/2.5/). All images appearing in this work are property of the respective copyright owners, and are not released into the Creative Commons. The respective owners reserve all rights.

This book was produced with Pressbooks (<https://pressbooks.com>) and rendered with Prince.

CONTENTS

Introduction to Beats Empire	1
PART I. PLAYLIST 1: BEATS EMPIRE EP	
1. Beats Empire, the Game	11
2. Formative Design Research	19
3. Assessment Games Versus Learning Games	45
PART II. PLAYLIST 2: BEATS EMPIRE LIVE	
4. Using an Evidence-Centered Design Process for Educational Game Design	65
5. Learning Analytics in an Assessment Game	83
6. Validating the Assessment using Cognitive Interviews	99
7. Participatory Design for Developing a Formative Assessment Dashboard	121
8. Designing Bridging Activities	145
9. Classroom Implementation	163
PART III. PLAYLIST 3: BEATS EMPIRE RETROSPECTIVE	
10. Beats Empire Hits and Misses	177

11. An Interview with Game Designer Luke Jayapalan	205
12. Concluding Thoughts	217
References	233
About the ETC Press	245

INTRODUCTION TO BEATS EMPIRE

SUMMARY

In this chapter we introduce our game, Beats Empire. We also discuss the purpose of assessment in education and how Beats Empire addresses certain long-standing barriers that assessment can enforce and at times accentuate.

Most of us have a story about a test gone wrong – something we should have aced but failed, some evaluation that did not capture what we knew to be true. Similarly, many teachers understand that what they learn from an assessment can unfairly represent the students’ knowledge and capabilities. At times, teachers and students recognize that this incongruence and misrepresentation is not a failure of pedagogy, understanding, or retention – but of the assessment itself. This likely contributes to the general classification of assessments not as “fun” or “engaging”, but instead, as a cause of stress.

As a personal example, when Matthew Berland (one of the authors of this book) was in sixth grade, he failed a test. He had not failed a test before – despite (or perhaps because of) having ADHD and anxiety; but he failed this particular math test. He

knew the processes and answers for every single question on the test, but due to an unfortunate situation involving another student, both ended up in the principal's office – and Matthew received an F on the exam. When Matthew got home, he locked himself in a closet and cried because – although this test predated “high stakes testing” – it felt very “high stakes” to him. According to Matthew, the most salient thing about that traumatic day, is his mother – a child psychologist at the time – consoling him through the closet door. She explained that the test was mostly a means to keep control of the class, and that the teacher already knew that Matthew understood fraction addition. She assured him that, despite the F, he would get an A in the class (true) since he knew the material (also true). Many children – even those not raised by psychologists – know that tests often do not accurately measure what they or their peers know. And some of these students, along with their teachers, understand that in the wrong hands these rankings can serve as mechanisms of control, rather than evaluations of understanding.

While assessments have a long history, the movement for standardized assessments accelerated meaningfully during the time that education became mandatory. There was an increase in students attending higher education (Mislevy, 1993), which in turn increased the need to make decisions about acceptance and placement. Often multiple choice or short answer questions were used due to time constraints and ease of scoring. However, these assessments can (and have) been used to amplify inequalities (Lindblad, Pettersson, & Popkewitz, 2018). Take the Harvard Entrance Exam from July 1869 as an example. While presented as an assessment that would predict if students would do well at Harvard, the questions such as “Name the chief rivers of Ancient Gaul and Modern France.” and “What is the reason that when different powers of the same quantity are multiplied together their exponents are added?” focused mainly on memorization, or were subject to [mathematical] interpretation. One of the most

shocking and fascinating things about this test is that so much of the knowledge is useless and inert – “Give all Infinitives and Participles of abeo, ulcisor” – though we have long known that assessment structured this way demonstrates very little useful or predictive knowledge (viz. McLelland, 2017). In other words, to get into Harvard, prospective students had to prove that they retained massive stores of knowledge that they could and would never use except on these tests. Did this test exist to reinforce class differences? It certainly served that purpose. Did it exist to require students to spend untold time memorizing useless facts just to prove that they were willing to do so? It definitely required will power – but more importantly access to the necessary information and the time to commit it to memory. Exams thus evaluated commitment to idle time, and were used as a means of ranking, gatekeeping, and control with respect to class, race, and gender. The agency to explore and create was only doled out carefully in bits of math or expository writing. The test therefore carried with it the weight of a future value judgment, the possibility of attending an august institution, and the perpetuation of preexisting power structures.

Moving from sixth grade math test through old Harvard exams and into the present, consider the United States’ public education system. The crisis that this system faces will only worsen if we do not reframe assessment to make it more relevant, creative, adaptive, and connected. To address this crisis, there have been several movements in the assessment field. For example, Mislevy (1993) discusses a need to shift assessment from knowledge to more *conceptual understandings* and the measurement of *practices*. This need is re-iterated in the National Research Council’s book (2001) which expresses the concern that traditional assessments are not focused on the skills and abilities that are most meaningful for the students. This is in addition to the concern that these assessments are not providing useful information to teachers and students for improving instruction (NRC, 2001).

In fact, research has shown that assessments that only provide grades can have negative effects on student performance, self-efficacy, and motivation, particularly for low-achieving students (Andrade & Heritage, 2018).

More recent efforts have focused on the use of formative assessment, which is assessment designed to influence instructional decisions and provide feedback on students' strengths and challenges. When used well, formative assessment has indeed been shown to improve outcomes for students (Wiliam, 2018). However, teachers do not always have a deep understanding of how to design and use formative assessment, meaning that this tool is often under-utilized in the classroom.

In 2020 when schools across the world shifted to remote learning due to the COVID-19 pandemic, this lack of understanding about the value and purpose of assessments among teachers, administrators, and policy makers became increasingly apparent. Assessment shifted back to previous mentalities: some teachers and policies measuring primarily factual knowledge (using multiple choice items); an emphasis on seeking ways to ensure students could not cheat (Jankowski, 2020). Many assessment opportunities became overt tools for surveillance and control when live video auditing and locked screens were introduced. These lockdown assessments typified and perpetuated negative assessment standards while simultaneously renewing worldwide educators' search for new ideas.

The pandemic heightened what these educators already knew: teachers need tools and support in determining ways in which assessments can be *beneficial* in their classrooms. Conversely to being used to surveil or control, assessments that take advantage of current technology can provide teachers with insights into students' practices, such as their problem solving skills and their agency and ability to apply their knowledge. Assessments that

can be administered on-line also have the advantage that they can be used both in the classroom and remotely. The development of these tools along with support for teachers in how to use these tools can greatly increase and diversify student motivation and performance in the classroom.

As we set out to develop a game-based formative assessment, our design process thus incorporated fundamental questions such as, “How do we make assessment a tool for agency?” “What does it mean for assessment to afford agency?” “How can we assist teachers in integrating positive formative assessment practices into the classroom?” Even when formative assessment practices are in place, the formative assessments are not always fun or engaging for students – and are often stressful or anxiety-inducing instead. Students who are given tasks like exit tickets (questions they must answer before the end of a lesson) or classroom discussion prompts are still aware they are being assessed – albeit in a slightly lower-stakes manner. Research on stealth assessments (e.g., Shute et al., 2016) has explored ways to integrate assessment more seamlessly into activities students enjoy. We are jumping off from that work towards a critique of the assessment’s function: if a tool affords enjoyment, agency, or creativity in one realm, how might it continue to structure bias in another? The tool explored in this book attempts to actively counteract bias in ways that are both meaningful and enjoyable for students. We are clearly not the first to think about overlaps between play, assessment, and bias – indeed, there is an excellent body of literature on building playful assessment into “maker” activities explicitly designed for diverse student populations (e.g., Blikstein, 2013; Kim, Murai, and Chang, 2021; Lin et al., 2020; Lui et al., 2020). That said, there is relatively little literature on building open-ended videogame-based assessments, despite the apparent prevalence of such games.

In this volume, we – an interdisciplinary team of researchers – share insights from three years of intensive design research

around a game for formative assessment of computer science and data science skills. The context, grounded in New York City public schools and the needs of middle school students and teachers, set a challenging innovation agenda. Beats Empire, the game designed in response to these challenges, quickly garnered many awards and downloads. The game offers each student an opportunity to manage recording artists, shaping the parameters of their recordings in response to their analysis of song trends in an imaginary metropolis. The learning and assessment themes align to a national computer science framework; the topics of data collection, storage, visualization, and interpretation are key data science skills that are relevant across many subjects, such as math, science, and social studies. Game play maps realistically onto how today's music industry producers and managers use data to increase their artists' followers, listens, and sales. Overall, Beats Empire can bring dry learning standards to life, helping teachers and students see computer science skills as connected to academic subjects, as addressable in school, and relevant to life beyond the classroom.

This book follows the team through the initial design process, the development of the game, its implementation in middle school classrooms, and the research developed around game data, and ends with reflections on lessons learned and future directions.

One foundational element of the design process and our research was designing for minoritized groups as well as including the voices of those for whom the game is ultimately created: teachers and learners. Teacher and learner responses to the game demonstrated the real-world challenges that the game addresses, and ways to improve. This feedback also opened the door for several additional avenues of research – some of which we were able to pursue while others will require additional time and work. We explore bridging activities to be used in the classroom that facilitate learning with the game while encouraging students

to develop thought processes and knowledge that they can expand beyond Beats Empire and the classroom. Additionally, we outline the development and utility of a dashboard created in conjunction with teachers across varied disciplines. This dashboard empowers users to understand what the data from Beats Empire says about the students' progress and provides teachers with active responses and strategies.

The book continues with an examination of the connections between the game and the real world. We specifically explore how these connections facilitate conversations about the real world value of computing skills and support broader participation in computer science and data science. Looking forward, we reflect on how this game design could inform further research on the connections among assessment, authenticity, and broadening participation. And finally, in keeping with our goal of designing and deploying an assessment that affords agency, researchers reflect on what *they* learned while designing the game, studying its use in schools, analyzing the data they collected, and exploring alternative forms of assessment in real public school classrooms.

The findings that emerge are relevant to game designers, assessment developers, teachers, and researchers. Computer science and data science are critical topics for the future of learning, teaching, and assessment in our growing knowledge economy. This book offers research-based insights on how we can design games for assessment that advance teaching and learning on these important topics in New York City and nationally and is a timely resource for other researchers who are working on similar projects or are interested in doing similar work. Towards those ends, this book presents a set of works that use Beats Empire, a classroom assessment game, as an object-to-think-with (Papert, 1980; Holbert & Wilensky, 2019) that was designed – with varying levels of success – to restructure assessment as a tool to support agency.

Teachers will find a substantive working alternative model of assessment that they can deploy in their classrooms immediately. We suggest tradeoffs and attempt to be clear about what information they will and will not be able to get from our assessment model. School leaders will find a new way to look at how they can assess and may be assessed in the future. In our experience, administrators are often looking for books that offer practical, practicable alternatives for assessment. Education researchers, college instructors, and professors will find a detailed, theoretically rich description of our play-based model of assessment. Game designers and developers will find valuable information – including interviews and a post-mortem – about how they might develop assessment games.

In closing, it almost feels trite to emphasize how much the years 2020-2021 changed the landscape of education. There are few schools worldwide that look unchanged from 2019. This book uses a set of design, theoretical, and research perspectives to suggest ways technology can be used to think differently about the purpose and value of assessment. Our experiences of education during the Covid-19 pandemic only reinforced our belief that assessment as we know it should and will undergo dramatic changes over the next few years.

PART I.

PLAYLIST 1: BEATS EMPIRE EP

CHAPTER 1.

BEATS EMPIRE, THE GAME

SUMMARY:

In this chapter we provide a full description of Beats Empire. This description should serve as a reference point as you read through the book. However, in addition to reading about Beats Empire, we highly recommend readers play the game which can be found at <https://play.beatsempire.org!>

Throughout this book we will refer frequently to specific design features of Beats Empire. To situate these discussions, it seems reasonable to start the book off with a detailed description of the game to give the reader a broad sense of the game's look, feel, and mechanics. While individual chapters will generally provide some description of the key game feature being discussed, it may be useful to bookmark this section for easy retrieval when needed. Our hope is that Beats Empire can be an exemplar of a playful assessment, serving as a model for future assessment and educational game design.

Beats Empire is a single-player game about music. Players take

on the role of a music studio executive and their goal is to use data about listener interests to make decisions about what artists to sign to their burgeoning label, what kind of songs to record, and where to release these songs. To win the game, players can either aim to release three number one hits in one genre, or release a handful of top five songs in multiple genres.

Beats Empire fits into the “management game” genre. As such, gameplay involves managing a series of decisions for the music studio. Possible decisions include signing artists, recording songs, researching ways to improve song quality (i.e., “buffs”), and releasing songs. Players enter “rooms” in the music studio where each decision is managed (Figure 1). These decisions are enacted primarily by spending two in-game currencies: money and fans. Once the player has completed their actions for the day, players progress the time forward one week to see the result of the decisions they have made.



Figure 1. The studio screen shows multiple rooms that players can visit to enact decisions. When they have completed their decisions they can increment time forward by clicking the “Next Week” button.

When the game starts, a tutorial directs the player to first hire artists. After clicking on the “artists” room players see a small list of possible artists to sign, some of which are individuals

and others that are full bands (Figure 2). While many artists are loosely based on real world musicians (for example Beyoncé was designed to evoke the artist Beyoncé), others are entirely fictional. New artists randomly become available to the player most weeks. Each artist is associated with one specific music genre and has a collection of available song moods or topics they can record. All artists also are assigned a specific value for a host of “songwriting skills.” These skills include ambition (the artist is more likely try for bigger hits at the cost of a higher chance for big flops), reliability (artists are less likely to make mistakes), talent (which indicates how quickly their songs can increase in quality), and persistence (how long they spend in the music studio working to record a hit). The higher the value (up to five) the better the artist is at each skill. Players can also give artists additional mood or topics for recording and can improve songwriting skills by spending money to upgrade the artist.

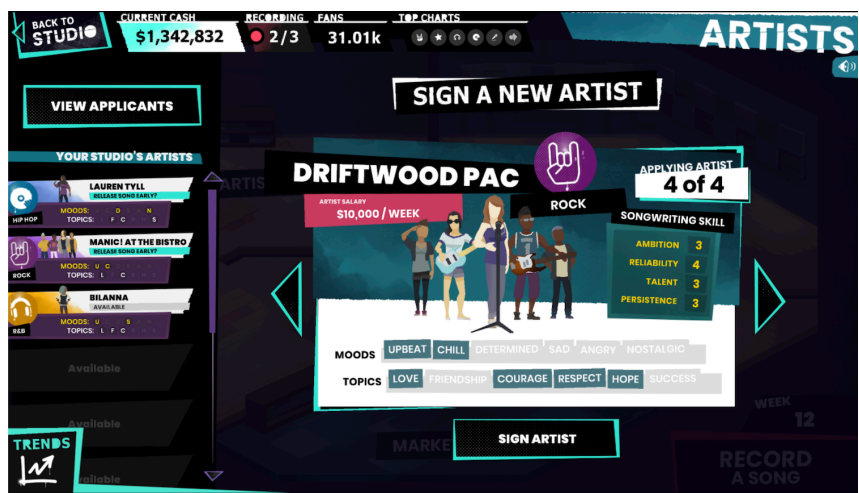


Figure 2. The artist signing screen allows players to use their accumulated money to sign artists. Each artist is assigned a specific music genre and can record a limited number of song moods or topics.

Once the player has signed at least one artist they can begin recording songs by clicking on the “recording” room (Figure 3). After selecting an artist, the player needs to decide which

borough in the fictional city they will target the song to, and which mood and topic the song will be about. Players can also optionally generate a song title. To make these decisions, players are encouraged to look at listener interests by clicking the “Find Trend” button. Doing so takes the player to the trends screen (Figure 4).

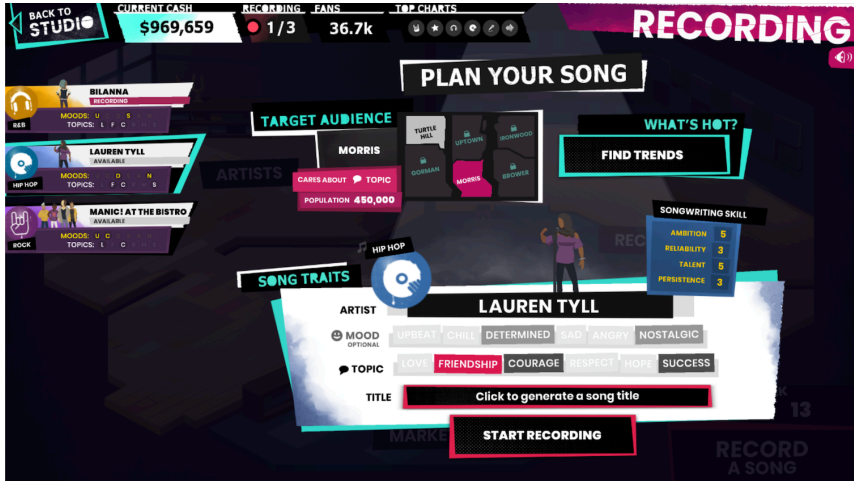


Figure 3. In the recording screen, the player has to select the song mood and topic as well as determine the target audience.

In the trends screen, players can view listener interests in song mood, topics, and genres in each city borough (Figure 4). This data can be viewed as a bar graph, line graph, or chloroplast. Each possible city borough has different interests so depending on which borough is selected, different data is displayed. Boroughs available at the beginning of the game are only interested in one song feature (such as topic), while boroughs that can be opened later may be interested in two or more song features. Likewise, this data is generated uniquely for each player and each game session so no two playthroughs of Beats Empire are alike. Moods, topics, or genres that the artist can record are highlighted. Players can click on a bar or line displayed to select a song feature. For example, in Figure 4 the player has clicked on the line for the topic “hope.” Doing so asks the player to

make a prediction. Here the player can indicate if they believe this particular topic, mood, or genre is the “most popular” or “trending up.” In addition to indicating their understanding of the graph feature, if they are correct this choice will also provide additional money or fans at the song’s release. When they are done, players click “Plan Song” to return to the recording screen where they can make any additional choices about the song to be recorded before they eventually click the “Start Recording” button. While songs are being recorded, the background music of the game shifts to a muted clip of the song the player has just chosen to record.

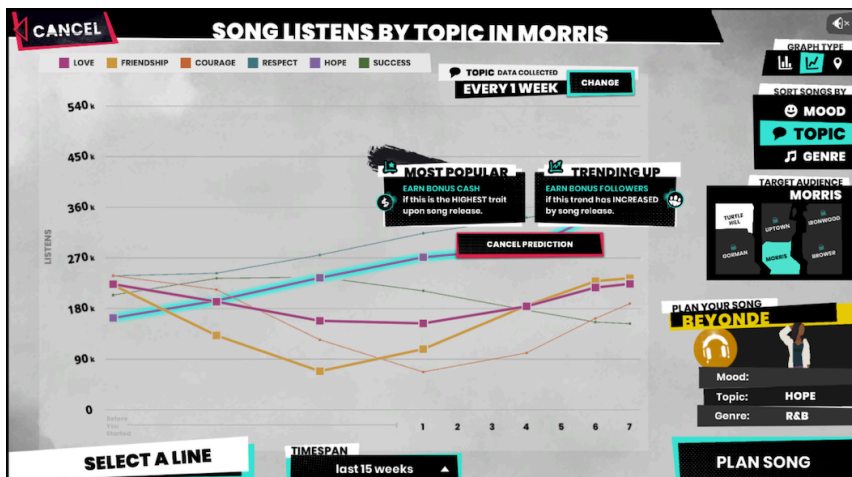


Figure 4. The trends screen allows players to view data on listener interest. Players can also make “predictions” about whether a song feature is the most popular in a borough or if it is trending up.

Songs generally take a few in-game weeks to record. Players can release a song early or wait until it is completed which gives the song a bit more time to increase in quality. To release a song, the player goes to the recording screen. Here the player sees a list of the features of the recorded song and an “awesome meter” indicating the relative quality of the song (Figure 5). After releasing the song an audio clip of their completed song plays. The specific song the player hears depends on the mood and genre of the song they have recorded with each combination

resulting in a different sound. Likewise, songs with a higher song quality score include additional sound layers meant to represent a more fleshed out song. This means that for every song released, players are likely to hear a different song, with “awesome” songs having a more full and produced sound. As the audio of the new song plays, a screen shows the player the result of their recent release, indicating the amount of money made by the song, the number of new fans attracted, and whether or not their prediction (if made) was correct. Finally, the player sees a “top charts” list which indicates how far up the week’s charts the song reached.



Figure 5. The song release screen indicates the details of the completed song as well as the relative quality of the song, indicated by the “Awesome Meter.”

While the trend screen provides a useful look at listener interests, to make accurate predictions players may find data collection occurring every three weeks inadequate. Data collection can be viewed and modified from the collection and storage screen. Here players can increase the frequency of collection for listener interests for topic, mood, and genre (Figure 6). However, collecting more data requires that the player increase their storage space to hold this additional data, which in turn requires

an additional weekly cost. Players must manage the tradeoff of better data versus the cost of storage.

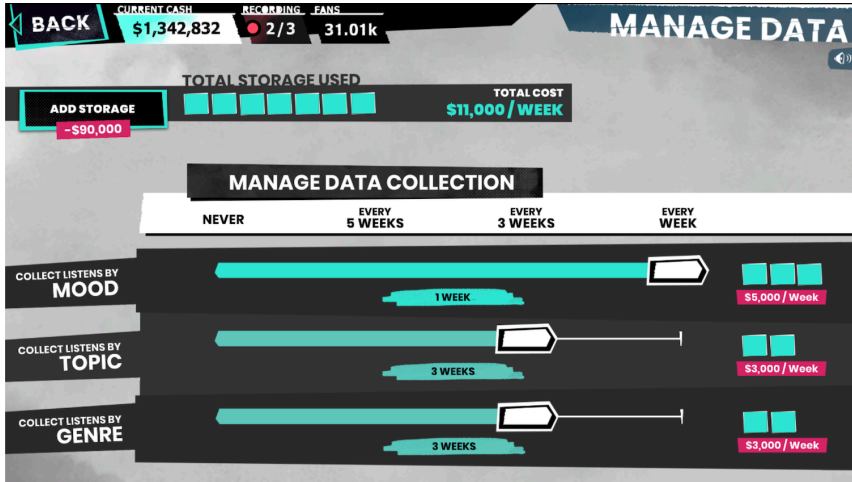


Figure 6. In the collection and storage screen, players can increase the frequency of data collection for listeners' interests in mood, topic, and genre, and can purchase the additional storage required to hold this data.

Finally, in the “market research” screen players can upgrade their recording studio and unlock additional boroughs to target song releases (Figure 7). Unlocking boroughs costs money, but each unlocked borough also allows players to target songs to additional song features and offers a large number of possible new fans, making it easier to release a song that rises up the top charts. Genre upgrades also cost the studio money but make finding good artists, earning money, and releasing top songs easier.

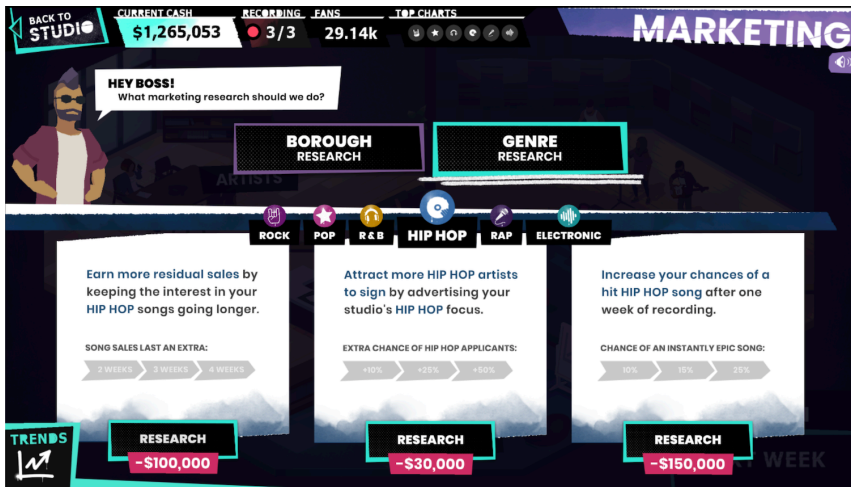
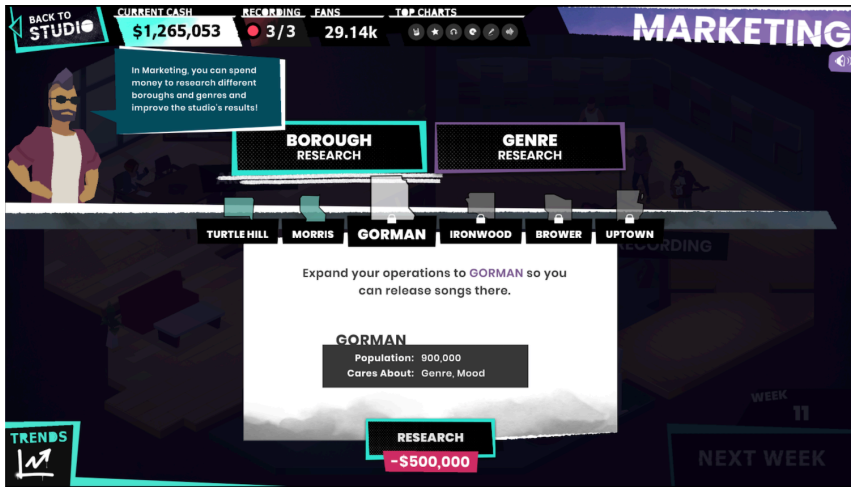


Figure 7. In the market research screen, players can unlock additional boroughs for song releases and acquire “buffs” for artists and song recording.

While the above description should provide a useful reference, the best way to truly understand the game is to play it! Beats Empire can be accessed at play.beatsempire.org. In addition to providing some useful context for subsequent chapters, hopefully you’ll also agree Beats Empire is fun!

CHAPTER 2.

FORMATIVE DESIGN RESEARCH

SUMMARY

This chapter discusses the formative work that went into conceptualizing Beats Empire. It serves two purposes, first to identify how we came up with the tenets of our design. Second, it serves as an illustration of how formative research, participatory design, and user-centered design can inform educational technology. To these ends, the participants in the development of Beats Empire ranged from computer science education advocates and educational administrators to teachers and students.

INTRODUCTION

We have all seen K-12 educational games fail. In some cases, this failure is because they are not engaging for the player, in other cases, they fail because educators do not see the value of introducing games to their students. It is easy for those of us in game design to assume we understand what kids want to play – after all, we were once kids who liked to play games. It is also easy for educational designers to assume what teachers might

use in classrooms – we are the “experts” in education. In reality, however, these assumptions are unreliable and lead us to design for our own experiences rather than the interest and needs of students and teachers. Fortunately, the design process can be used to check those assumptions – enabling us to counter false assumptions and confirm useful ones.

While introductory engineering textbooks often start with “needs assessments” to figure out what people need, much design in education simply presumes knowing what people need without asking them. This is rooted in something real and useful: we are teaching students and we are expected to know what they should be taught. We do not ask students to decide whether “ $2+2=4$ ” is useful, and we should not privilege, say, folk theories of epidemiology over the scientific knowledge of the epidemiology community.

That said, when a group of older people steeped in careers far from the setting of K-12 schools start deciding “what kids relate to” without talking to any students or “what teachers need” without working with teachers, it is a foregone conclusion that the project will miss the mark. Therefore, before we started developing Beats Empire, we conducted a series of research efforts allowing the students and teachers to design and comment on elements of an educational gaming experience, helping us to better understand how to design the game for diverse classrooms. It was crucially important to talk to a diverse set of teachers and students with a wide range of values, needs, and preferences.

As this chapter will show, this turned out to be tricky but, we believe, worth the effort. We knew early that it would be complicated when PI DiSalvo, an expert in participatory design for minoritized youth, correctly suspected that the initial design proposed by our external game designers reflected the musical influences and experiences of the design team (rock bands at

small venues) – rather than the musical preferences of NYC youth. So she ran quick design research activities with students in the target demographic. Out of all of the students who responded, exactly one listed that they listened to “Rock” – many (many) more listened to, say, K-Pop, and none of them went to small venues (i.e. bars) to hear bands. The one student who listened to Rock was such an anomaly that we (again, informally) asked the student why she listened to Led Zeppelin and Jimi Hendrix. Her answer was simple: she listened to “Classic Rock” to (lovingly) annoy her father, a well-known local Hip-Hop artist. At the risk of sounding very old – teens are teens!

That said, we did not begin the process without any design insights: the research team came to the Beats Empire project with years of design research with youth from minoritized communities¹. This experience informed our initial design, prompting us to focus on a meta-design structure in which youth could bring their own interests and identities into the learning experience. And we met that goal: Beats Empire is fun. It speaks to a wide range of youths’ interests and makes real-world use of data relevant to young people.

Because Beats Empire is fun, we often receive positive feedback, along the lines of “what good idea that was” or “how lucky to hit upon something that kids like”. However, it is important to recognize that the concept for Beats Empire did not just appear in a moment of inspiration. Instead, the design can be traced back to research done on the rollout of computer science (CS) education in New York City (Holbert, et al., 2020), a careful review of the types of computing curriculum offered, and the National Computer Science K-12 Framework (K–12 Computer Science Framework, 2016). Additionally, it is based on a previously established meta-design framework that

1. Minoritized communities refers to neighborhoods, organizations, or groups that have traditionally been seen as disenfranchised members of society and have limited access to the tools and processes in computing. (Crooks, 2019)

conscientiously centers young people's diverse cultural influences and identities (DiSalvo & DesPortes, 2017). In order to ensure interest and relevance, after we developed the initial concept of a music studio, we worked closely with teachers and students, bringing them into the design process using Participatory Design. This collaborative design process additionally included experienced game designers at Filament Games who were similarly dedicated to infusing these concepts of relevance into each aspect of Beats Empire. (See Chapter 11.)

In this chapter, we will briefly outline the research that led to the game concept. We then trace two research activities that helped us flesh out the game design. First, a participatory design activity with teachers, and second, the design research with students that was noted above. While there were other design research activities, these two highlights the range of engagement one can have with stakeholders in designing educational games.

Educational game design is different from traditional game design, because while creating a fun game was a goal, it was equally important to meet educational objectives, appeal to students from diverse backgrounds, and create an activity that teachers would be willing to use in their classrooms. We know that deep knowledge of the disciplinary content – in this case, computational data and analysis, and a deep understanding of the educational context – from the district level to the individual classrooms, are both necessary to design educational games. Our use of design research is an acknowledgment that while we had the disciplinary knowledge, we needed the participation of students and teachers to design for context.

FORMATIVE RESEARCH FOR CONCEPT

The initial seeds for Beats Empire came from interviews and observations with administrators and teachers in the NYC Department of Education (DOE), academic papers, news articles,

social media posts, and educational documents related to the rollout of CS education in NYC (Holbert, et al., 2020). One day during this research phase, Holbert, Berland, and DiSalvo were in NYC conducting a long interview session with Michael Preston, the Director of CS4All, Debbie Marcus, the Executive Director of CS Education for NYC DOE, and Leigh Ann DeLyser, a lead researcher for CS4All. In this session, they sketched out the learning ecosystem in NYC schools for CS education. These key players described the impetus of implementing CS education in the NYC school curriculum as needed by industry efforts and their philanthropic arms. These industry leaders in turn influenced NYC politicians and eventually brought on CS education researchers and advocates to implement CS education in the schools. The groups that were brought in last were the school administrators, teachers, parents, and students...those most impacted by the change. This meant that the implementation of concrete learning experiences were floundering behind the policy and publicity of the CS4All efforts.

Following this meeting, the research team sat in a coffee shop, revisiting the interviews and consolidating notes. At this moment, we identified how the funders, CS education researchers, administration, and non-profit organizations were a significant presence in conversation about CS education in NYC and seemed to be the driving forces for what was being implemented in the schools. Perhaps because of this focus on higher-level policy, rather than classroom implementation we noted that a critical piece was missing from this learning ecology: the assessment of CS learning. As we sat in that coffee shop, the idea of creating a tool for assessment seemed like a difficult (and uninspiring) task. But as the team began thinking about how assessment could be exciting to us, we focused on ideas around creating playful assessment – something that students would find fun rather than repetitive or dull, and something that teachers would prefer to implement than yet another quiz.

Playful Formative Assessment

We conducted participatory design sessions and informational interviews with people working at for-profit, non-profit, and higher education organizations all involved in developing and implementing CS curriculum and professional development for teachers. Through the course of this research, we identified several needs for CS assessment tools in the schools. Of particular interest to us and fitting our skills in developing playful and culturally contextualized tools, was the idea of providing a formative assessment that could assist teachers who had little CS training. This resulted from our finding that while a great deal of effort was being put into creating a CS curriculum, many teachers with little CS background had a difficult time assessing what worked and what did not work to reach the learning goals (Holbert et al., 2020). We, therefore, focused our design efforts on creating playful formative learning assessments. Initially, we were imagining quick interventions that could provide teachers with some clear measures to see if their efforts to teach CS were working, how they might be off track, and which students needed attention. Our brainstorming ranged from collectible cards that represented core concepts to “bug” finding adventures that helped with identifying common errors in computer programming. But before focusing on the type of play that the assessment might provide, we looked into the learning goals we might assess.

Data and Analysis for Middle Schools

There are four strands of curricula in the CS K-12 Framework: Computing Systems, Networks and the Internet, Algorithms and Programming, Impacts of Computing (K-12 Computer Science Framework, 2016). We knew that in order to focus our project we needed to choose just one of these strands. In talking with Leigh Ann DeLyser, one of the leaders in CS4All, she identified the Data and Analysis strand for middle school students as a ripe

opportunity for formative assessment because many of the Data and Analysis learning goals were already tied to the current Math and Science curriculum. Despite this connection, Dr. DeLyser pointed out that transferring data and assessment knowledge from one context (e.g., math, science) to the CS context was something that teachers with little CS background found difficult. They found it even more difficult to identify if they were meeting the learning goals for the CS K-12 Framework. This discovery was supported by the experience of each researcher: DiSalvo's informal interviews with many non-CS teachers in NYC about another computing project, Nathan Holbert's networks as a faculty member at Teachers College in NYC, and Matthew Berland's years of experience in designing CS learning technology for the classroom.

Meta-Design Framework for Diverse Cultural Context

Our team produced a number of ideas for the game context, but a driving objective was a game that would appeal to the remarkably diverse student population of the NYC school district. We chose to use DiSalvo and DesPortes' (DiSalvo & DesPortes, 2017) work on designing project-based learning experiences that scaffold students' design process to bring their own values and interest. While we brainstormed ideas across a broad range of contexts, such as game design, music composition, visual design, collectible cards, and creatures, etc. we eventually landed on music as a central piece in cultural identification for youth.

Within the team there was debate about leveraging music composition because it is rich with opportunities for exploring algorithms and music production. The meta-design framework, however, helped us focus on music production, which allowed students to identify with the genre of music they chose and also to set their own goals, such as making money, critical success, or dominating a neighborhood or genre. In contrast, music

composition would likely only appeal to those who had an interest in making music, not just listening. We hoped that this meta-design approach to learning – creating a learning environment that allows students to design some of the context and experience – was a way to create culturally sustaining pedagogy that could be used in diverse classroom settings as each student would be able to customize and design their experience to fit their interests and values (Paris, 2012).

REFINING WITH PARTICIPATORY DESIGN AND USER-CENTERED DESIGN

We saw the potential to create something new with educational technology. Namely, we saw a chance to create a playful formative assessment rather than relying on traditional assessment. We thought that playful assessments could be used more frequently with students and had the potential to help teachers who had less CS expertise understand what CS concepts their students had mastered and which ones they needed to work on. We also saw promise in aligning that idea with middle school learning goals from the Data and Analysis strand of the CS K-12 Framework. We had brainstormed ideas that fit that criteria and focused on developing a simulation game that used data and analysis skills in creating and managing a music recording company.

However, we had not talked with teachers to see how they might use playful assessment in their classrooms and we had not talked with students about the ways they understood the music business, what music they listened to, and what types of game goals might motivate them. At this stage, we wanted to engage these stakeholders – teachers and students – in design research, as they would hold us accountable to what served them best. We chose to use both participatory design and user-centered design approaches. Participatory Design, from the Scandinavian tradition, is often misinterpreted by educational researchers and

learning scientists as User-Centered Design (DiSalvo, et al., 2017). The distinction between these two designs is not simple (Frauenberger, et al., 2015), and for this reason, the use of these two terms in this chapter is defined below.

User-Centered Design is a broad term that denotes that the design process is informed by the end-users (Abrams, et al., 2004). This usually means putting an artifact, such as a similar tool or a low-fidelity prototype, in front of a potential user in order to test the usability of an artifact or ascertain what they like and what they would change about the design. The goal for user-centered design is to refine choices that designers have already articulated or specific elements of an artifact or learning experience. In user-centered design, the people one talks to are giving direct feedback on a design artifact, not sharing personal stories or reflections. Because of this focus, typically user-centered design is not considered human subjects research and does not require institutional approval to conduct. However, it also limits what one can report about participants, typically not sharing details and only reporting findings in terms of general outcomes or design implications rather than individuals' input.

Participatory design is related to user-centered design but is also a continual practice and a democratic philosophy (DiSalvo, 2014; Hansen, 2019). A critical component to participatory design is to scaffold the design process for individuals or groups with the goal that they fully participate in a collaborative design session. These scaffolded design activities should prompt exploration of the design context and encourage reflection on the use or non-use of artifacts. These activities should also provide participants with the building blocks to create designs by building on their expertise and their lived experiences – without previous expertise in design, subject matter, or pedagogy. Because participatory design provides room for reflection on how designs fit into participants' broader life, including subject matter that might put the user at risk if shared, we consider this

human subjects research. Because of this, we seek institutional review of our participatory design protocols to make sure we are keeping subjects safe and holding ourselves accountable to the highest standards of human subject research. This process allows us to share more details about our findings with minimal risk to participants.

For the Beats Empire game design we leveraged both user-centered design and participatory design.

PARTICIPATORY DESIGN WITH TEACHERS

Our design process began with participatory design activities with middle school teachers. The overall goal was to prompt them to reflect on both how they might currently be teaching data and analysis in their class and how they might formatively assess learning. We met in small groups with NYC public middle school teachers who represented a range of content areas (Literature, Math, CS, Science, Art, Music and Health) and teaching experience (1 year to 39 years). Teachers were guided through a set of design activities in which they discussed the relevance of the data and analysis strand to their current instruction and reflected on the use of digital games for assessment in class. We hoped that the teachers would contribute to the game design as well as an understanding of the broader ecology of the classroom and formative assessments they might use so we sought approval to conduct research on our interactions with teachers from review boards at our institutions and NYC schools.

We started the design workshop by explaining our goal to help NYC schools better integrate the CS K-12 Framework into their existing classes. All of the teachers were aware that their schools and administrators wanted them to integrate CS concepts into their existing classes in addition to offering new courses focused on CS. Seven out of the eleven teachers had taken at least one

professional development workshop to learn about integrating CS into the class. All of the teachers, including the CS teacher, recognized that this integration was not an easy task. However, some were still very excited to try, while others were resistant to add to their class loads.

Activity 1: Designing Assessment for Data & Analysis Learning Goals

The design activities consisted of three parts: the first was to uncover whether any of the concepts from the Data and Analysis strand were already being taught in their classes; the second was for teachers to design formative assessments that they might use to evaluate learning progress for those concepts; and third to get their ideas on how they would incorporate an existing game that relied heavily on data. We started by giving the teachers short descriptions of the concepts in the data and analysis strand, which included:

1. Collection of Data
2. Storage of Data
3. Visualization and Transformation of Data
4. Inference and Modeling of Data

After we gave the teachers these descriptions they asked us questions about the concepts. While data collection, visualization, inference and modeling were concepts the teachers were familiar with, they had little grounding in how data was stored and why that mattered and they did not understand how or why data was transformed for computational use. Teachers asked us questions to clarify what some of the concepts were. For example when asked why data storage was a hard concept we gave a concrete example — showing how things like colors were transformed and stored using RGB numeric representation. Teachers were then asked to complete a design worksheet to think through the activities they currently use to teach Data and

Analysis concepts or what they might envision teaching and how they might assess students' learning of those concepts (Figure 1). We encouraged them to be creative in how they might integrate Data and Analysis into their classrooms. Teachers completed between 1 and 5 of these worksheets with most providing 2 – 3 classroom activities. When they completed the description we specifically asked them to add stickers that might be representative of the assessments they would use for such activities, some pre-printed and other's blank for them to fill out. Finally, we asked them to include post-it notes that highlighted the data and analysis concepts that might be learned from the lesson.

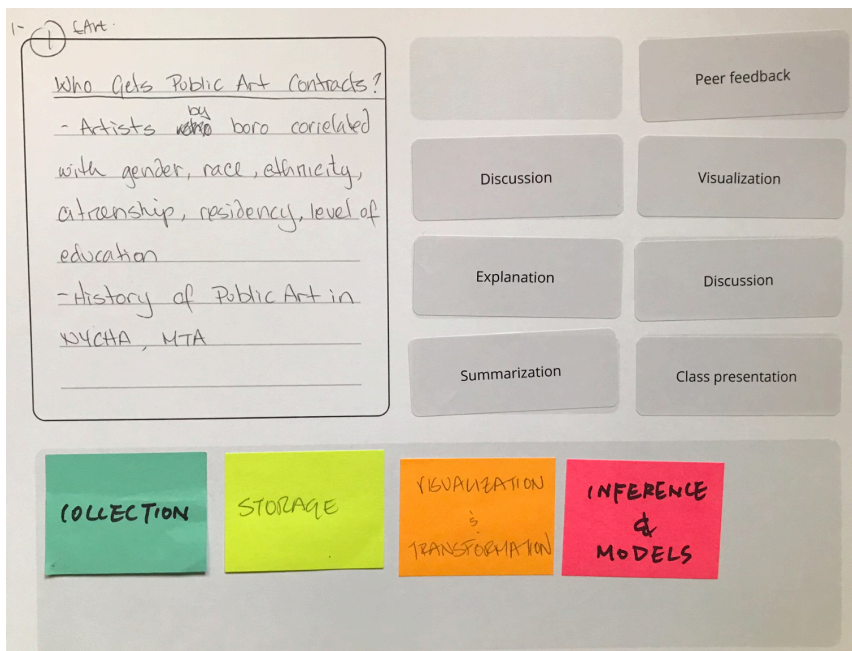


Figure 1. In the participatory design teacher worksheet, teachers first filled out the open area by describing how data and analysis is currently taught in their class or by brainstorming new ideas for teaching data and analysis. Then teachers added stickers with methods of assessment and finally post-it notes that corresponded to the learning goals. In this example an Art teacher suggested creating a visualization of data on who gets public art contracts. This would help students think critically about how art is funded and what “Public Art” means in addition to helping them learn about data.

What we learned from the activity. To analyze the data from this activity we reviewed the audio recordings of the sessions and our notes. Authors Gupta and Zhou created bios of each teacher, addressing their background, the types of lessons they designed, their hesitations and enthusiasm for the idea of including data and analysis skill development in their class. Building off of these bios we developed three fictional personas (Pruitt & Grudin, 2003) that highlighted important and actionable items for designing data and analysis lessons and assessment (Figure 3).

Art, Music, and Health: To our surprise, the Art, Music, and Health teachers were the most enthusiastic about integrating CS learning goals into their curriculum. The health teacher was already working with the computer science instructor/coordinator to develop a project that incorporated data from a health class survey into both spreadsheets and Python coding projects. The two art teachers brainstormed the most ideas and tied data tracking to classroom behavior, geographic data on local hip hop artists, and creating a visualization of who receives, and who doesn't receive, art funding. The music teacher generated an idea very similar to Beat Empire. What we learned was that these teachers feel their subjects are valuable, but the emphasis and funding being allocated to STEM content in schools often left them behind. They hoped that integrating CS content would afford them more status or funding but recognized they would need to work with CS experts to do this right.

History and English: In contrast, the History and English teachers were hesitant if not openly opposed to incorporating CS content into their classes. The history teacher was vocal about his reluctance to incorporate any aspect of computing into his class. He explained that his priority was making sure the students were literate. He was concerned that computers would obfuscate the real learning goals for history he was trying to teach. He spoke passionately about how many of his students couldn't read for

content or understand the context of what they were reading and he felt that history was one of the few classes in middle school that would teach those important lessons. The English teacher had an idea that focused on argumentative writing with a little inclusion of data and analytics. Similar to the History teacher, the English teachers (quite rightly) felt their primary goal, to get kids reading literacy rates up high enough to pass standardized tests and engage in the world, was difficult enough without introducing CS content.

Math and Science: We had assumed that the math and science teachers would be very receptive to the inclusion of data and analysis concepts. However, their enthusiasm was muted. Similar to the history and English teachers we expect they had standards to teach to and too little time to achieve them. However, when we helped to identify how the CS concepts related to standard math and science concepts they began brainstorming. Their ideas lacked some of the creativity of the art, music, and health teachers. This suggests that math and science teachers might need hand-holding to create interesting data and analysis projects and help create a transfer from math or science to computer science (see Figure 2.).

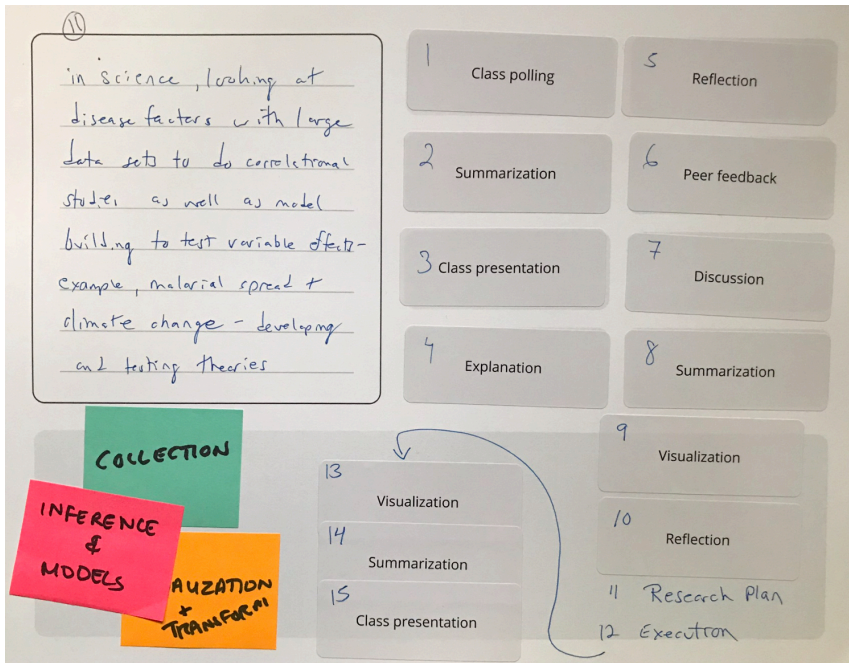


Figure 2. A science teacher sketched out a data project connected to disease factors. They saw that current learning goals for their class overlapped with CS data and analysis thread.

CS Teachers: The teachers with the most expertise in CS represented a wide range of teaching experience, with one having over 20 years of teaching and the other in their first year. This range highlighted how expertise in CS is only one aspect of integrating CS into other disciplines. The novice teacher had very compelling ways to make computing more relevant, including one idea about tying in online bullying to a project-based learning experience that would look at data storage and privacy and then use computational tools to construct a story or narrative about a dramatic event that would get kids interested. However, she did not participate in conversations that might have helped other teachers find connections. In contrast, the more experienced teacher was already working with the health teacher (see above) and was excited about integrating computing

into other courses. She wanted to know more about the art teachers' ideas and to see if she could build on that interest.

The worksheet scaffolded teachers in creating a classroom lesson on the fly about a topic that they may not be familiar with. While the teachers were experts in creating the activities for their classroom, the step-by-step process of reflecting on assessment and returning to the learning goals for the Data and Analysis strand caused the teachers to consider and communicate issues and design inspirations related to these two issues in particular. Without this activity, we may have missed the exciting art, music, and health connection to CS topics. We also noted how strong personalities can infect others. Negative personalities can reduce participation (history) and the enthusiasm of the Art teachers encouraged others. This highlights how participatory design is a group process and it is important to facilitate conversation and collaboration between participants.

Activity 2: Designing Assessment for Data Game

In a second activity, teachers watched a video walk-through of a simulation game being played and answered questions. With this activity we sought to understand how much they currently used games, if they used (or could imagine) games as a way to assess learning.

We started the activity by asking the teachers to share their previous experiences with games – digital or physical – in the classroom. We then asked them more specifically about their familiarity with simulation games such as SimCity (SimCity Games) or Rollercoaster Tycoon (Rollercoaster Tycoon).

We then showed a video of a person playing a game called Plague Inc. (Plague Inc.). Plague Inc is a simulation game where the player embodies a type of bacteria/virus, and then has to infect and kill all the humans in the world before the humans can develop a cure. Before watching the game we asked the teachers

to take notes in order to think about whether the player's learning has achieved goals that we have just discussed from the Data and Analysis learning goal list. After viewing the game, teachers were asked to provide feedback on how similar games might fit in with their lesson plans and what game data they would find useful for assessment. We had the teachers use the cards with the names of each data and analysis learning goal and asked them to talk out loud about how they would assess these learning goals based upon the gameplay. We specifically asked them about assessing individual students and the classroom as a whole, and what they might change in the game design to better understand if students were learning.

The video was short, and teachers appeared engaged when watching. They took a few notes and looked at a few screen captures as a group. When giving their reflections they tended to let one teacher lead and then offer small suggestions to the assessment plan. In retrospect, we realize that the activity might have been better if we had first watched the video as a group, then given them a worksheet with paper copies of all images to write their reflections on, then finally asked them to share out.

The history teacher, who at first had been very reluctant to teach any data content liked the game, mentioned that kids would enjoy the dark theme and that he could do something with it (although he did not elaborate). Other teachers started reflecting on practical issues with having students play a game like this that stood apart from other content they were teaching. Two suggested that this type of game would be good for "exit tickets", an activity that students do to fill the last 10 min of class. They noted that there were lots of graphs but were unsure how that might be teaching data and analysis content. They also noted that it would be difficult for them to interpret how much kids were learning from playing the game and they would need some sort of reporting mechanism to understand learning outcomes.


Design implications. As a research practice, the participatory design activities with teachers highlighted a number of important implications for our game design process. First, it's important to facilitate conversation and collaboration between participants during activities so they can support one another. And be prepared to have many examples available to prompt and inspire participants during these conversations.

Second, strong personalities can infect others and negative personalities can reduce participation. The History and English teachers were not receptive to teaching any CS in the classrooms as they felt they already had too many standards to teach. These attitudes impacted the degree to which others in this session felt comfortable sharing openly.

Finally, be open to unexpected insights. As we assumed, the Math, Science, and CS teachers were receptive and saw how this game-based assessment of data literacy could fit into their curriculum. What we did not expect was for the Health, Music, and Art teachers to be so enthusiastic about integrating data and analysis into their classes. Furthermore, these teachers were willing to integrate game-based assessment into their classrooms for content in which they did not have expertise. We may have missed the exciting art, music, and health connection without this activity.

Based on all the activities we constructed three persona, imaginary teachers that represented many of the issues and opportunities that arose during the participatory design session (Figure 3).

SECONDARY PERSONA



Rachel

7th grade Mathematics teacher

Previously taught Middle school and college level Math

5 years experience teaching

Not taken CS courses

STORY

Rachel is a Math teacher for the 7th grade in a public middle school. Her subject has a specific curriculum with pre-determined topics and units and she has to make sure she finishes the curriculum in time and her teaching meet the state standards. To teach, she employs creative methods like hands-on projects and game type activities in class. Sometimes students use Google softwares on chromebooks to do these activities, first to do individual work, then do group discussion and reflection after that. In these projects, she is already teaching some of the CS K-12 data and analysis concepts but she does not tie them to the CS framework. She is excited to adopt CS learning, but is not sure how it would fit into the curriculum.

PAIN POINTS

- Harder to take risks in class content since had to meet standards and cover syllabus
- Her time is limited

NEEDS

- A formative tool to test CS skills
- Help in making that connection of how this tool could be incorporated with what and how she currently teaches

WANTS

- Meet state standards and cover the curriculum
- Understand students learning progression formatively
- Teach and assess students so they truly understand the content

Figure 3. This example teacher persona synthesized mathematics and science teachers' perspectives for the game design and dashboard design teams.

Our design takeaways from these sessions included:

1. A game focused on learning about data has potential for cross-disciplinary use but will fit better with some subjects than others. Subjects that have direct overlap with learning about data analytics, such as CS, Math, and Science will be an easier fit while CS lessons that cross over with Health, Music, or Art will necessitate more support.
2. The amount of time the game takes to learn and play needs to be shorter and more flexible so that teachers can better fit this assessment in with all of the other material they cover.
3. Teachers will need a great deal of guidance in understanding data that is produced by an assessment game, therefore a teacher dashboard will be a critical part of the design. (See Chapter 7.)

USER-CENTERED DESIGN WITH STUDENTS

Once we had a feasible design for a data assessment game from

the teachers' perspective, we needed to determine how to connect our concept to students' interests and aesthetics. This was highlighted when the game design firm, Filament, presented an early iteration of the game which focused on 5-member bands and venues where they would perform. Our experience working with youth in our target audience made us push back on that suggestion for a couple of reasons. First, middle school kids rarely attend music events – their knowledge of music was gained through word of mouth, social media, and music apps. Additionally, we felt that music – particularly in the world of Hip Hop and Rap – was no longer centered around independent bands, but instead collaborations between artists, producers, DJs, and others. As Filament revisited and adjusted their design, we checked in with middle school students to fill in details about what kinds of music they liked and what they thought about music production. We also wanted to understand how they reacted to educational games in the classroom. Because this was research to directly inform design, rather than human subjects research, we only gathered data on the design feedback and the reporting is limited (as explained above).

Methods

Unlike the Participatory Design with teachers, we used more traditional user-centered design research methods with students, meaning that we did not ask the students to design things or play design games (Ehn, 2008). By continually checking design directions with our audience, however, we hoped to keep the philosophy of participatory design intact in our work. We conducted two focus groups with a total of twenty-two students in an after-school session at NYC middle schools. As students came into the room they saw two activity stations. The students self-organized with most of the boys in one group and most of the girls in the other. The two groups took turns participating in both activities. Because this activity was focused on the design of a product, and not on research, we did not record the students

during these events and only report on the design implications rather than the student comments and reactions.

The first activity allowed the students to play a music production simulation game for about 15 minutes. It had no educational intentions but was built only for fun play. After the activity, the students talked about their experiences with the game: what they liked and what they did not like.

The second activity was a more traditional focus group. We asked the students a series of questions about what games they liked to play, how they felt about educational games, what kind of music they liked, and what they knew about music production. Finally, we showed them a color printout of an initial mock-up of the studio in our proposed game to gain their insights about what they envisioned about the aesthetic and conceptual model of a music studio.

Design Implications

These sessions helped us in a number of ways. First, they let us know that they were enthusiastic about a music production simulation game and liked playing games in school whenever they could. They also confirmed our ideas about how students thought about music production. They talked about musicians as artists, rather than bands. And used cloud-based services to find, share and purchase music. They found new music from suggestions from friends and influencers both online and in person, such as people handing out CDs on the street. Generally, they were not reading magazines or going to concerts – music for them was online entertainment. Regarding the aesthetics of the game, the students liked the styling of the characters’ however, they found the initial mock-up a little confusing and wanted to center the musical instruments in the layout, making the vibe of the game to be more about the music.

We then communicated these implications to Filament, relaying

more details about what students were interested in through personas that captured the different types of students we saw playing the game (Figure 4). These personas are imaginary students that capture a number of different characteristics we identified during the activities.



Michael

8th grade student at a public middle school in Bronx, NYC

MUSIC INTERESTS

Favorite Music:

Hip Hop, Rap, R&B

Favorite Artists:

Little Uzi, 21 savage, Migos

Dislike for:

Country, Jazz, Love songs

LISTENS TO MUSIC

Technology

Headphones
DJ System that is connected to his phone
Bluetooth speakers

When

While going to school
Anytime he gets an opportunity (almost all day)
When playing video games on his Xbox

FINDS NEW MUSIC

Hears it from friends
Hears it on the radio 97.1
Hears it when someone drives by on the street
YouTube/SoundCloud/Spotify/Pandora/Tunes autoplay
Shares playlists with friends

GAMES

Likes Call of Duty, 2K Basketball, GTA, Mortal Combat, Rainbow 6, Fortnite. Multiplayer games. Competitive one-on-one games. Plays for having fun with his friends. Plays for "lots of hours." Most of his free time.
Dislikes games with story. Doesn't want to have to read to find out things.

THOUGHTS ON MUSIC PRODUCTION

His friend goes to the studio, "edits their beats and rolls it out to soundcloud," Michael believes that "That's how most of the artists we know make music is on SoundCloud. Start out with a wack setup..." His friend uses a laptop. Once his friend gets a popular song, people start to follow him. He stays active and keeps uploading songs so people don't drop off.

Michael thinks that artists write one song, then during down time they start writing the next song. Also, they talk to people to find out what they think about it and they get influence from other rappers. A rapper goes to producers and publishers and other rappers to find out if they should put the song out.

TECHNOLOGY USE

Personal devices:

Laptops, computer at homes, gaming systems, owns a Samsung smartphone. Uses them for social media, text, youtube, some school work—google drive.

At school:

Desktops/chromebooks/macbook airs



Michael

8th grade student at a public middle school in Bronx, NYC

MUSIC INTERESTS

Favorite Music:

Hip Hop, Rap, R&B

Favorite Artists:

Little Uzi, 21 savage, Migos

Dislike for:

Country, Jazz, Love songs

THOUGHTS ON EDUCATION GAMES

Do you like any education games?

Math game we played at school.
Minecraft—teaches you architecture. I feel like it's good. They're ok.

Which education games do you like?

"Run" in Math class. Challenge games. X moves to do something. "Color switch",
Catch a bird and then do a problem

Which ones don't you like?

Don't move a character, just type and type. You can see the character or you are the character. Shoot a bird with the right answer—don't like those.

STORY

Michael is an 8th-grade student in a public school in NYC. He is considered a popular kid and has a lot of friends. He plays a lot of games with his friends. After school everyone (all the guys) play together. Sometimes this is at one person's house but more often it is online. Music and games are tightly linked for Michael. He and his friends share music while playing games and find new music from games. His brother is a DJ so he knows about producing music and the music industry. He is interested in getting into DJing and is building his own set up at home. He knows you have to "Spend money to make money," and is willing to invest in new equipment.

Figure 4. This example student persona synthesized one of three student perspectives to help the game design team better understand the audience.

CONCLUSION

It is difficult to design good games. Games that serve a specific purpose, such as assessing computer science, can be even more difficult. This project sought to bring top-down design directives from administrators in NYC schools and the CS K-12 Framework architects in line with what students wanted and teachers needed, without ignoring our own design inclinations. The top-down directions had us focus on CS assessment rather than creating a learning game and our own perspectives on learning pushed us to create something playful – to think about games as a playful way to assess. The students’ values and interests not only sparked the focus on music but directly shaped the way we portrayed the music industry in the game. Finally, the teachers helped to design by giving us practical feedback on how game-based assessment might work in the classroom and pushed us to invest more effort into a dashboard for teacher feedback.

The teacher design sessions also helped us understand how to ensure that formative assessment needs would be met. They did this first by providing us with information about which subject areas this game might align with. We chose to focus on developing the game as an assessment tool for Math and Science courses while exploring the potential for creating a data and analysis learning game for Music, Health, and Art classes. We also recognized that gameplay needed to fit into the class time, so a 30-minute experience with the game would be the longest we could expect a teacher to use in a 45 min class, and a shorter 10-minute experience would likely be used more often by teachers. Finally, we recognized that teachers liked the idea of using games in the classroom, but understanding what learning had occurred was a challenge. Because of this, we invested time into designing a teacher dashboard to provide actionable formative assessment to teachers. (See Chapter 7.)

We also centered the students’ preferences and needs by

checking in with middle school students to make sure the concept was appealing, to fill in design details, and to understand how they reacted to educational games in the classroom. They let us know that they were enthusiastic about a music production simulation game and playing games in school whenever they could. We learned about what kinds of music they liked, and they reframed musicians as artists, rather than bands, confirming that they used cloud-based services to find, share and purchase music. When asked about the aesthetics of the game, the students responded that while they liked the styling of the characters, they found the initial mock-up a little confusing. All of this information was used to create fictional personas for middle school student users.

The personas of both the teachers and students that might use and play *Beats Empire* that we created after these sessions became critical artifacts to communicate to the game design and development teams. They helped to shape the many small design choices, such as the aesthetics and music choices of the game and helped to shape the general game mechanics, pushing for more of a music management mechanic rather than a singular band perspective. In our interview with the designers in Chapter 11, the game designers remarked on how the persona's helped them design for the correct audience. Based upon the student personas the game designers moved forward with music production, rather than a band management model. These personas were also reflected in the choices for the music genres: Rap, Hip Hop, Pop, electronica, and Rock, and impacted the names, which were playfully modified from contemporary artists such as Beyoncé, Envision Gryphons, and Micki Mirage. This provided a more authentic representation of middle school children's musical interests and identities.

In designing learning and assessment games, there is a unique challenge beyond making a great game that will attract players. We needed to make a game that filled a need for teachers, that

could be used in the classroom, and that resonated with a broad population of students. While great game designers can make great games on their own, the extra burden that learning and assessment games need to meet, make design research necessary to produce a fun and productive game experience.

CHAPTER 3.

ASSESSMENT GAMES VERSUS LEARNING GAMES

SUMMARY

Learning scientists and educational game designers have developed a host of principles and theories for building effective learning games. However, creating a useful assessment game requires a different set of priorities and considerations. In this chapter we explore four design tensions our team encountered in the development of a game-based formative assessment. These tensions included the design of structured versus open game play, narrow or broad coverage of target content, how to treat player choices that conflict with assessment goals, and how to build in player progression. In this chapter we discuss the implications of these tensions and detail the choices made by our team, and their implications, in the design of Beats Empire.

INTRODUCTION

When gathering a team to create Beats Empire we intentionally involved experts in both learning game design as well as experts

in assessment design. However, we found early on in the design process that combining assessment and game design is no easy task!

Games provide immersive and interactive spaces that invite players to problem-solve, explore complex ideas, try on new identities, etc. (Gee, 2003). Well-designed game-based environments situate learners in stimulating contexts and encourage players to explore compelling ideas in personally meaningful ways. Two decades of work by educators and designers have shown that video games can therefore be powerful spaces for learners to engage with commonly taught ideas, practices, and concepts (Steinkuehler & Squire, 2014).

At the same time, a renewed classroom focus on supporting students' engagement with practices – rather than attending solely to content knowledge – presents a challenge for building effective assessments. When building assessments around content knowledge one simply has to look for the presence or absence of that knowledge; multiple choice and fill in the blank exercises are an example of this kind of assessment. However, if we want students to know how to *do things*, for example, reason through complex ideas, engage in argumentation with evidence, etc., then assessments must capture this thinking and action as a process over time. Likewise, practices are not isolated bits of information, rather they unfold in context. Consequently, when assessments fail to meaningfully engage learners – when students complete decontextualized assessments simply because they've been told to – it raises questions regarding the validity of claims made from these assessments. Still, teachers need effective methods of determining what their students know and can do. This has led to a new interest in games as assessment tools (e.g., Kim & Miklasz, 2021), as games can provide opportunities for students to apply their knowledge to interesting and complex situations that unfold over time in a game and can help students

demonstrate how they would use knowledge and skills beyond the game.

No doubt learning games often include assessment moments. However, we have found that designing a game to measure what students have learned in another context is quite different from checking to see that students have learned something while playing a game. In other words, designing an assessment game is quite different from designing a learning game! In this chapter we explore four tensions between designing a video game to be a learning environment versus the creation of a game-based formative assessment.

Key Considerations When Designing Learning Games and Assessment Games

Approaches to the design of educational games vary widely. While there are many possible ways to characterize educational game design, for the sake of brevity we will simply examine game designs that differ in the degree to which “content” – or what is to be learned from the game – is integrated into core game mechanics and interactions. Some games separate the game-like activity from the aspects of the game meant to promote learning. A classic example of this design is of course *MathBlaster*, where players solve simple math problems between bouts of blasting aliens. While modern games have improved, making the “learning” phases less starkly separate from the “playing” phases of the game, many educational games continue to include concepts, practices, and skills as things to be learned through screens of text rather than in the playful phases of the game. In contrast, some educational games integrate the content to be explored in core game mechanics. In these games the way you play the game IS the thing to be learned. For example in the game *DragonBox* players balance a field of abstract symbols as a way of exploring key principles of algebra. Here, game mechanics

embed key features of the learning goal, making gameplay itself a necessary component of the learning process.

Regardless of where a game falls on this integration spectrum (sometimes referred to as *conceptual integration*, see Habgood & Ainsworth, 2011), playing the game well means improving the ability at using the game system to do interesting things. Consequently, effective educational game design requires that learners make sense of the “system” of the game. *In learning games, learning goals and game design goals are aligned.* Designers create experiences that help players understand the system which, in well integrated learning games, aligns in some way with systems, phenomena, concepts, practices, etc. that have meaning beyond the game.

Assessment is the art and science of collecting evidence that allows us to draw conclusions about what learners know or are able to do (Pellegrino, 2014). There are many different types of assessment and many different ways to divide the assessment space. From formative versus summative assessments, to classroom versus large scale assessments. In embedded or stealth assessments students are not necessarily aware that they are being measured (Shute et al, 2016) – as opposed to more traditional assessments in which they are fully aware of the significance of their situation.

As we will discuss further in Chapter 4, assessment design starts by specifying what we want to know about students, and subsequently developing tasks that provide students with opportunities to demonstrate their knowledge. This process of specifying what we as educators and game designers want to know about learners and identifying how they might formulate this knowledge should always be based on a theory or model of cognition that defines the nature of knowledge – what it means to “know” something – and the process of “learning.” A useful cognitive model allows one to map out the important aspects of

the construct to be learned, describes ways in which a learner would engage with that construct, and identifies how a learner might represent and/or demonstrate their knowledge.

There are many modern theories of cognition, and each has implications for the design of learning and assessment environments (Plass, Homer, & Kinzer, 2015). *Cognitivism* has been the primary model of learning in the American classroom. In this model of cognition there is an assumption that knowledge is constructed by the individual through their experiences in the world, and that this knowledge exists “in the head” and is carried with us as we go about our day from context to context. Think about how this works in a school: students are meant to learn an abstract idea in the classroom, and then “transfer” or apply it in concrete and new situations out in the world (Haskell, 2000). Returning to game design, if we adopt a cognitivist model in the design of an assessment game, then we might aim to create experiences that invite players to demonstrate in the game that they have acquired and can use some knowledge learned in another context, such as a classroom.

Alternatively, many learning scientists now recognize that knowledge is not isolated from the world in which that knowledge is learned and used – that knowledge is in fact *situated, social, and cultural*. In these *situated* models of learning knowledge does not exist as an individual mental abstraction, but rather is *always* an interaction between the individual and the environment, the social and cultural context, the body, prior experiences, etc (Greeno, 2006). A situated perspective suggests that “knowing” is always “knowing with.” Treating knowledge as situated, social, and cultural has massive implications for assessment design! One cannot remove or hide the context in which knowledge is learned or applied because there is always a context. Rather, assessment requires inviting the learner to put knowledge in action – to use the context for thinking and reasoning.

Connecting the Game and Classroom to the Real World

Adopting a situated perspective blurs the distinction between a game for learning and an assessment game. In particular we note that games for learning and games for assessment both should leverage interesting and relevant narrative elements and enable players to meaningfully interact with target learning or assessment goals. While in some cases a game's aesthetics may be thought to be separate from the game's mechanics, we find that the integration of the two impacts not only whether someone wants to play the game, but also how they engage with game mechanics, and what knowledge they use for thinking and reasoning both in and beyond the game (for some examples of this see Hunicke et al., 2004).

For *Beats Empire* we believed it was important to identify a game genre that would appeal broadly to New York City youth. This matters for both learning and assessment games because neither can be successful if players believe the game is unrelated to their interests or if they are unable to activate and leverage their existing knowledge – much of which is situated in their awareness of “how things work.” Using the participatory design techniques described in Chapter 2, we found that music is widely appealing to students, forms a core part of their interactions with one another, and is something they believe they know a lot about. For a learning game, this experience with music means students could use their intuition and prior knowledge about music as they make sense of complex game mechanics. Likewise, for an assessment game, players' interest in music would ensure they were motivated to interact with game elements and to demonstrate their expertise.

And yet, for students to connect game experiences to contexts outside of the game (including both classroom and “real world” contexts) it's important that game mechanics and representations have a reasonable alignment with the target

content in the real world (Holbert & Wilensky, 2014; 2019). That is, while the game may still include fantasy elements, the player should be able to explicitly note connections between the game world and the real world beyond it. In Beats Empire, we invite players to see a connection between the game world and that of data science through the use and analysis of a variety of data representations – in a context they are familiar with. Players make inferences about trends in listeners’ changing musical preferences by viewing line and bar graphs. These representations are core to game play and assessment. By evaluating which representations players choose to use, when and how players use these representations, and the inference they draw from these representations we are able to build a model of their understanding of data analysis practices.

DESIGN TENSIONS

In the above sections we have identified important overlaps in the principles of design for learning versus assessment games. In particular we note the importance of adhering to a coherent theory of cognition and of building connections between gameplay and the reasoning and use of knowledge beyond the game. However, there are a number of choices that can be made in the ultimate structure of such a game that can lead to a variety of tensions between mechanics and features that would ultimately result in a game that is optimal for learning, for assessment, or just plain fun. In the design of Beats Empire we found four tensions. The discussions among our team that emerged from these tensions generated food for thought surrounding the values and priorities central to the design of learning and assessment games.

Structured Versus Open Gameplay

One of the largest tensions we encountered when developing Beats Empire was the degree to which the game is highly structured – directing all game actions for all players down a

narrow and uniform path, versus open – where each player has the freedom to choose how they will interact with the game.

In a learning game where target constructs are embedded in the core game mechanics, all gameplay exposes players to learning opportunities. Consequently, whether the game is highly structured or open, game designers can be sure players encounter the learning constructs of interest. The same is not true for assessment games. For an assessment game, navigating this tension of openness versus structured gameplay requires reflecting on the purpose of the assessment, and how the information from the assessment will be used.

An assessment must collect evidence of knowledge in a way that is clear and accessible to all learners. In an assessment game, designers must be intentional about how and when players will encounter opportunities for assessment. It is also important that *all* players have the chance to demonstrate their knowledge. If players are not provided the opportunity to engage with the constructs being assessed, then the assessment is not able to collect evidence about each student's knowledge! Assessment games (and assessments in games) typically address these requirements in two ways: by making the assessment moments overt such as with pop ups or after-level quizzes, and by highly structuring gameplay thereby ensuring that all players pass through the assessment opportunity.

But heavily structured games with few opportunities for unique decisions and actions can limit players' personal investment in a game. Similarly, pop ups or after-level quizzes are an impediment for players to do what they really want to do: play! These design features persist, despite decades of research on learning design suggesting that students learn more when they are personally invested in what they are learning (Papert, 1980; Holbert, Berland, & Kafai, 2020). Consequently, to ensure that players have the opportunity to meaningfully engage with

content and/or assessment moments, it is also important for the game to provide opportunities for players to play in a way that makes sense to them.

In Beats Empire we chose to address this tension by making the gameplay somewhat open. We did this by allowing players to make decisions about which artist to sign, which songs to record, where to release songs, etc. We wanted students to be empowered to build their own unique music empire rather than simply moving through the steps of a plan predetermined by the game designers. This choice came with tradeoffs.

If players are free to focus on any particular genre in Beats Empire, or record a song with any combination of features, how will we know if the choices they are making reflect their understanding of the data, or simply personal preference? The short answer is we do not always know! For example, because graphs of listener interests are a core part of gameplay we do know students encounter – and consequently have opportunities to reason with and learn from – these important data representations. However, because players are not required to make explicit inferences around specific uses of these data representations, it is possible players may spend many minutes in the game looking at these line graphs without ever indicating whether or not they understand how to draw an inference from the graph.

Had Beats Empire been designed to be a summative assessment game, where the aim would be to determine whether or not a student has passed a certain threshold of competency on a topic, the lack of structured opportunities for assessment would be detrimental. However, as a *formative assessment game*, where the goal is to gain information that will assist in the teaching process, the tradeoff between openness and structure can be more balanced. All student choices about how to interact in the game provide information to the teacher. By documenting the context

in which players' make choices, and designing game mechanics to make these choices more explicit, we can remove some of the guesswork about why a player made a particular choice.

For example, in *Beats Empire* students can choose to view data about listener interests using a bar or line graph. Knowing which data representations each student prefers, when they choose to view each, and how they progress through the game with these representations can provide valuable contextual information to the teachers. Using information about what students know and can do with these data representations teachers can consider how classroom practice may need to adapt in response. Likewise, for *Beats Empire* we developed a “prediction” mechanic in the game where, after selecting a specific genre, mood, or topic for a song to be recorded, players can indicate whether they believe this graph shows the feature to currently be the most popular among listeners, or a features that is “trending up.” This player prediction is essentially an opportunity for the player to indicate their interpretation of the graph. If the prediction is correct, the player receives a bonus of in-game currencies. Players are not required to make these predictions, but by making them easily available throughout the game, and providing some in-game reason for doing so (in this case, providing bonus in-game currencies), we create another opportunity for players to make their thinking visible to teachers.

Narrow Versus Broad Coverage

Another tension is within the game's scope. We grappled with the extent to which one game can assess a large variety of different concepts with limited development time and budget. As we note in Chapter 2, *Beats Empire* was built to assess computing concepts and practices in the K-12 CS Education Framework's Data and Analysis strand for middle school learners. This strand includes the topics: Collection, Storage, Visualization and Transformation, and Inference and Models. When designing

Beats Empire, core game mechanics were focused on the Inference and Models topic, and to a lesser extent, Visualization and Transformation. For example, when players make decisions about the features of a song to be recorded, they look at various graph types of listener interests and draw inferences about what features (genre, mood, topic) are likely to make for a popular song (Inference and Models). Players do not transform this data themselves, but they do have opportunities to look at and draw inferences from multiple representations of the data (Visualization and Transformation).

On the other hand, despite many ideas, debates, and prototypes, the team struggled to meaningfully integrate the Collection and Storage topics into the game design. Ideas for integration included having players create surveys for listeners, managing company data storage and file types, etc. While each of these ideas *would* integrate some aspects of Collection and Storage into the game, none easily fit the main gameplay loop we had settled on and all felt distant from the game studio narrative we had created.

Ultimately we did include a set of game mechanics that touch on Collection and Storage (though they are less integrated into the main gameplay loop than the mechanics that include Visualization and Transformation, and Inference and Models). When viewing graphs of listener interests, players can view a Collection and Storage screen where they can change the frequency of data sampling for each song characteristic. If the player chooses to sample listener interest more frequently, listener interest graphs include more data points, but the studio must in turn buy more storage space to hold this data. However, many of our own play tests and data from classroom implementations of the game suggest that experimenting with these game mechanics is not necessary for in-game success. Players can progress through the game just fine using the default settings for Collection and Storage. Consequently, without

explicit direction to explore this game mechanic we find that few players use or even find the Collection and Storage screen.

Ideally, Beats Empire would include a more effective integration of the Collection and Storage topics. However, each concept or practice that is covered and assessed in the game needs development time, and consequently more money, in order to ensure that it is covered in a way that is meaningful to both the students and the teachers who are using the information from the game. At some point the team had to recognize that not everything could be included. In Beats Empire this meant reducing the depth of the game mechanics that address the Collection and Storage concepts from what we had originally hoped. This alteration, however, allowed us to focus more resources and energy on the development and refinement of game elements that integrate the Inference and Models, and Visualization and Transformation concepts.

What Does a “Wrong” Choice Mean?

We faced the third tension in the design of Beats Empire when considering what it means for students to get things wrong in the game. In an assessment, errors are generally assumed to be evidence of an incomplete understanding of the target concept. However, an error may also indicate a misunderstanding of the question, task, etc. In other words, a student may *know* the target concept or practice, but may fail to demonstrate that knowledge. The potential for this misunderstanding is heightened in an assessment game context where the actions and systems players interact with are complex and players may have a diverse range of motives for each action. So when designing an assessment game, there is a real challenge in determining whether errors made by the player are due to an incomplete understanding of the target concept/practice, a failure to understand how to play the game, or actions motivated by something other than those

assumed by the assessment (we'll return to this topic again in Chapters 4 & 5).

Efforts to design a game that would make it possible to distinguish between a player demonstrating genuine knowledge or true misunderstanding frequently led to tension between designing complex versus simple game systems. This tension emerged early in the design process as the team considered the primary mechanics players would use to interact with data. Some ideas involved having players collect data on listener interests themselves, rather than be provided with this data; others focused on having players construct representations of data using an in-game querying language and graphing tool. While these ideas would integrate key concepts and practices such as Data Collection and Visualizations and Transformations into the game, they would also require extensive tutorials to “teach” the player how to use these complex systems. Likewise, complex systems can often lead to unexpected or odd results. For example, what if a player tries to query and graph data they haven't yet collected—would we just show a blank graph?

From an assessment perspective, a player failing to generate a usable graph may indicate they do not understand the necessary transformation and inputs to translate a data table to a graph. This would be a meaningful result! However, from a gameplay perspective, we cannot be sure whether the mistake is due to the lack of knowledge about graphing or due to a failure to understand this novel in-game querying language. This difference is crucial. After all, not understanding how to use a new game mechanic is not the same as not understanding how to construct a useful graph of data. Additionally, game systems that are difficult to use and have the potential to fail frequently aren't very fun.

Eventually the idea of having players collect and query their own data was abandoned in favor of providing players with premade

data sets. For each game played (from start to end), a unique data set is algorithmically generated so that no two players, and no two sessions, encounter the same data. This design is simple to explain to players and flexible enough that they can tinker and experiment with interpreting the data without worrying about getting lost in the game system. While this design decision means the game lacks more authentic data collection and graph construction opportunities, it allows players to focus on choosing from and making inferences about and with existing data representations.

We additionally found that while a simple game system addresses the challenge of determining if player errors are due to lack of content knowledge or a misunderstanding of game mechanics, players may still act on data representations for any number of reasons. For facts-based topics – such as determining which genre of music displayed on a graph has more listeners – determining if the player is right or wrong is straightforward. But evaluating whether the player has made a thoughtful choice using data is more difficult. For example, how do we know that players chose to record a rap song because of how they interpreted the listener interest graph, rather than simply because they really like rap music? Because we opted for a relatively open design for the game (see the previous section) it is not always clear if students' in-game decisions are driven by their interpretation of data or by personal interest. In such a game what does a *wrong* choice even mean?

One way to navigate these competing interpretations of player actions is to consider how the game goals might communicate what counts as “correct” play. When gameplay is scored using a points-based system, every in-game action is inevitably scored as correct or incorrect. Rather than focus on a score, or money or fans (two points-based resources that are present in the game), the goal of Beats Empire is to produce quality records. To win the game players can either focus on one particular genre and work

to get three number one hits in that genre, or aim to produce songs that break into the top five for multiple genres. Setting the game goal in this manner does two things: it encourages students to personalize the design of their music studio, and it ensures that in-game success is based on a broader gameplay strategy, rather than correct or incorrect individual choices. Instead of Beats Empire being a series of successes or failures, play is centered on learning about the relationship between data representations, decisions made about artists and song characteristics, and the relative success of each released song.

Progression Without Levels

Finally – and related to the discussions surrounding success in the game – the design team frequently debated methods for representing player progression. Levels are a game feature that are commonly used to gradually increase the game difficulty or as a means of representing progression to the player. For a learning game, attaining new levels communicates to the player that they are gaining knowledge or ability as they play the game. In an assessment game, passing levels may indicate the achievement of some benchmark. However, levels are a coarse metric, at best, and too readily used as a shorthand to measure students' understanding. While this can indeed be useful for a quick take, it is problematic if a student's progress on game levels is used in isolation – for example students that have passed three levels are automatically deemed more knowledgeable than students that have only passed two. As a formative assessment tool, we were therefore overly conscious about not just *what* Beats Empire would measure, but also *how* teachers would make sense of, and use, the data presented about students' gameplay. (See Chapters 6 & 7.)

While the lack of levels is common in many management games or other sandbox game genres, it is important that gameplay stays meaningful even after students have played for greater

lengths of time. Consequently, even in a management game it is important to provide a sense of progress and discovery. For example, in *Beats Empire*, while players can use accumulated money and fans to train more artists, they can also use these in-game currencies to conduct “market research,” which unlocks new song recording upgrades and opens additional in-game locations to release songs. Upgrades (a mechanic common in role-playing games) provide players with a clear sense of progression—their studio is becoming more *powerful* in that it is easier to record bigger hits that attract more fans. Similarly, having players unlock new boroughs in the fictional game city requires players to manage more song characteristics at recording, simultaneously gradually increasing the game difficulty in a way that fits the game narrative.

Progression in a game can be represented in many ways and we encourage assessment game designers to think more broadly than game levels or stars representing how well a player has completed a level. Progression in an assessment game should not just represent thresholds of skill or knowledge, but rather evidence of the player’s ability to use knowledge and enact practices in context.

CONCLUSION

While these four tensions may not be the only ones that arise during game design, they are significant aspects to consider when balancing assessment design with game design. Games are never just learning games or just assessment games. They are always both. Designing an effective assessment game that meets the needs of both teachers and students requires the design team to acknowledge their theory of cognition, and to clarify the purpose of the game and the critical aspects that would allow students to show proficiency. It also requires creating understandable game systems and effective supports that

students of all levels might need while playing the game but that do not hinder players as they attempt to show their proficiency.

As we note above, designing an assessment game also involves more than just designing the assessment. The design team must consider how to communicate the results of the assessment to the teacher. Designers must consider what information the teacher will see, and how they might interpret and act on this information. This awareness integrally impacts design choices throughout the game, including the tensions we explore in this chapter.

One unintended result of dealing with these game design tensions was a deeper appreciation for a diverse team who were willing to work collaboratively. As discussed by Luke Jayapalan, lead game designer, in Chapter 11, having experts in both assessment design and game design working together was vital to ensure that the resulting game provided valuable information to teachers and was also an engaging and meaningful activity for students. And while we did not always agree on whether a particular design feature met the goals we set out for the game, there was a strong recognition by both the learning and assessment designers that there would be compromises made throughout the development of the game. These discussions surrounding game design and game content provided a sense of clarity for the whole team regarding the overall purpose and usefulness of Beats Empire.

Finally, it is important to note that designing a meaningful and useful assessment game is hard work! And like most designs, we are not likely to get it exactly right the first time. We encourage designers to build frequent opportunities to communicate with stakeholders throughout the design and development cycle. For the Beats Empire team, this meant interviewing middle school students and teachers in the earliest brainstorming phases (as discussed in Chapter 2) as well as sharing works in progress

(including art, paper prototypes, early functional prototypes, etc) throughout the process.

Still, while we recognize that the version of Beats Empire widely available at the writing of this book could be further tuned and modified to strengthen its assessment characteristics, we are quite proud of what we have created. Beats Empire is a *fun* game! The game mechanics we designed as a team enable players to build personally interesting music studios that can represent not only their understanding of data and data representations, but also their love of music. And as we show in the subsequent chapters, we were able to draw a variety of conclusions about student learning, teacher practices, and assessment game design in the design, development, and implementation of Beats Empire.

PART II.

**PLAYLIST 2: BEATS EMPIRE
LIVE**

CHAPTER 4.

USING AN EVIDENCE-CENTERED DESIGN PROCESS FOR EDUCATIONAL GAME DESIGN

SUMMARY

Evidence-Centered Design (ECD) is a commonly used process for developing assessments which requires the developer to clearly define the assessment (and/or learning) targets, specify the evidence that is needed to measure students' ability related to these targets, and determine characteristics of tasks that would provide students the opportunities to produce this evidence. Working through the ECD process in conjunction with the game design process can ensure that the game will provide valuable feedback to teachers. In this chapter we discuss how the ECD process was used to work with game designers to develop a game used for supporting students and teachers in the classroom.

INTRODUCTION

As mentioned in previous chapters, we designed Beats Empire as a formative assessment game that provides actionable feedback

to teachers, meaning that it should provide guidance on what next steps teachers can take to best support their students in learning and engaging with Data and Analysis concepts and practices. One widely accepted process for assessment development is an Evidence-Centered Design (ECD) process (Mislevy & Haertel, 2006). While this process has most often been used to develop standard classroom and standardized assessments, it has also been applied to the development of game-based assessments (Mislevy, et al. 2014, Shute & Sun, 2019).

An ECD process focuses the assessment developer, or in this case the game designer, on decisions that need to be made to ensure that students—game players—are given the opportunity to provide evidence that, when evaluated, provides valuable feedback about the student. ECD provides a structure organized around what we want to know about students (also called the constructs of interest), what evidence students can produce that allows us to draw conclusions about the students, and how to develop tasks that allow students to provide that evidence. This process organizes the information needed to develop assessments around three models, the student model, the evidence model, and the task model. The student model defines what to measure or say about the students; the evidence model describes what behaviors and/or observations we need and how these observations can be interpreted; finally, the task model focuses on identifying task features that would allow us to make the desired observations.

As an example, let's say you're in a band and want to find a new guitar player for the band. To decide between the people who audition for the band, you create an assessment. First, you decide what would make a good guitar player for your band – maybe you prioritize being able to play a range of musical styles at an advanced level and care less how well they can move on stage, or vice versa. Having a clear definition of what makes a good

guitarist for your band helps you think about how to structure the audition process and specify the task. For example, if you are interested in certain types of chords or style of music, you may consider having each person who is auditioning play the same song, and the song you pick would cover the characteristics you are most interested in. Next, you have to decide what it is about their playing that is most important to you. Is it that the player hits each note exactly as intended? Is it that the player is clearly enjoying playing and is able to draw you into their song? Knowing what you are looking for can help ensure that you are evaluating each candidate in the same way and that the result of the evaluation will provide you with information you can use to make your decision about who should join the band. In this example, the student model was the ability to play guitar, defined by specifying the type of chords and music that you wanted to hear played; the task model was that the person would play a song with specific characteristics; and the evaluation model was the characteristics by which the candidates' song playing ability was judged.

Model	Purpose	Example	ECD in Beats Empire
Student	Defines what is to be measured or said about the students	Ability to play guitar or specific chords and types of music	Assessment targets Example: Ability to create a visualization for a dataset
Evidence	Describes what behaviors and/or observations should be encouraged	Characteristics on which playing ability was judged	Evidence required Example: The accuracy of a generated visualization
Task	Identifies task features that allow the desired observations to be made	Assign song that showcased necessary characteristics	Assessment guidelines Example: Students should be provided with a purpose and/or question to address and then either provided with or asked to generate a dataset or data visualization(s) to address the purpose or question

Table 1. ECD Process Models

While there are many different approaches to ECD, all approaches focus first on clearly defining what constructs need to be assessed and what it looks like for students (game players) to engage with these constructs. This information is used at all

stages of the development process and documentation is developed that describes the alignment between the features of the task and the constructs being measured. This helps ensure that (1) the resulting assessment is aligned to the purpose of the assessment and (2) there is documentation of the decisions that were made, which can then be used in the development of similar assessments or to explain the rationale behind design decisions.

This chapter will describe how our team worked through four phases:

1. Define assessment targets (student model)
2. Determine the evidence required (evidence model)
3. Develop assessment guidelines (task model)
4. Use these models to develop game features

While the narrative is presented linearly, the artifacts developed at earlier phases were often revisited and revised based on decisions made later. This iterative process established alignment between the phases and ensured that the resulting designs accurately reflected all of the decisions being made.

PHASE 1: DEFINE ASSESSMENT TARGETS

In the first phase of this process, we developed a set of assessment targets. Assessment targets are a set of knowledge and skills aligned to the standard or concept you hope to measure, and stated in a way that clarifies expectations of students who demonstrate those skills. A clear set of assessment targets provides guidance for what we can say about students based on how they perform on the assessment. In the example above for the guitar player, the assessment targets would be the chords that we wanted to know if a guitar player can play. Assessing for these chords would be targeted, which means you can only draw conclusions about the chords included in the

assessment (or audition), not every type of guitar playing. For this project, the assessment targets were developed by examining the K-12 CS framework for guidance on middle school data and analysis concepts. The assessment targets were grouped based on the K-12 CS framework specification of four components: Collection, Storage, Visualization & Transformation, and Inferences & Modeling (Table 2).

We found that there was a strong relationship between Collection and Storage components, as well as a relationship between Visualization & Transformation and Inference & Models components. These two relationships led us to define two sets of assessment targets that focused on these components separately and on how they can be combined. For example, data storage depends heavily on what data is collected, so creating an assessment target that has students consider both of these aspects at the same time (DCS3 in Table 2 below) highlights for teachers the importance of the integration of these skills.

During the development process of the assessment targets, we focused on what students can do as a reflection of what students might know. For example, our first assessment target listed (DCS1) focuses on identifying variables or data types that should be collected, which would likely require players to know what would constitute “data”.

In developing our assessment targets, we wanted to focus on two big ideas: the relationship between data that needs to be collected and the questions that data is being used to answer; and how conclusions are drawn from data. The targets listed in Table 2 represent our interpretation of the standards as related to these big ideas.

Data and Analysis Strand	Assessment Targets	
Collection and Storage	DCS1	Ability to identify variables or types of data that should be collected based on the purpose of the data collection
	DCS2	Ability to identify how to automate data collection (e.g., how often to collect the data, and the use of a computational tool to collect the data)
	DCS3	Ability to identify an appropriate representation for the data that is to be collected and stored given the purpose of the data and the storage constraints (includes identifying types of metadata that might be collected).
	DCS4	Ability to manage the trade-offs between data collection and storage requirements.
Visualization and Transformation and Inference and Models	DVT11	Ability to create a visualization for a dataset.
	DVT12	Ability to identify which data should be used to address a certain question. This includes identifying outliers and creating rules for the computer to filter outliers.
	DVT13	Ability to use the data to create an appropriate model that demonstrates relationships within the data.
	DVT14	Ability to interpret data models and visualizations for making predictions or drawing conclusions.

Table 2. Assessment targets for the K-12 CS framework's data and analysis concepts

PHASE 2: DETERMINE THE EVIDENCE REQUIRED

Once we developed our initial set of assessment targets, we then identified types of evidence to measure students' abilities related to these targets. For this evidence, we considered what actions students would do, what artifacts students would produce, and how these actions or artifacts could be evaluated (See Table 3). These actions were aligned to specific assessment targets to highlight what students would be expected to produce, and what quality of the product is important. This information is useful to developers as it guides how to structure tasks and what to focus on when generating rubrics or scoring guides for the tasks.

For example, looking at the ability to create a visualization for a

dataset (DVTI1), the evidence we would look for is the accuracy of the generated visualization. For a student to demonstrate this we would expect students to be able to create their own visualization and their creation would be evaluated by how well the visualization accurately reflects the data. Notice that this does not state anything about how well the visualization provides insight to answering a specific questions about the data. This latter ability is related to the ability to develop an appropriate visualization for answering a question (DVTI3), while DVTI1 focuses on how well students can represent given data in general. The requirements of the task and/or visualization should be further developed in the task model.

Assessment targets		Possible Evidence Used to Measure Targets	
DCS1	Ability to identify variables or types of data that should be collected based on the purpose of the data collection	E_DCS1	The appropriateness of students' choices in the variables they use and the types of representations they use to address a question
DCS2	Ability to identify how to automate data collection	E_DCS2	Whether or not the students' set-up related to automated data collection is sufficient for the question they are addressing
DCS3	Ability to identify an appropriate representation for the data that is to be collected and stored given the purpose of the data and the storage constraints	E_DCS3	The degree to which the representation of data the student is using is appropriate for the question they are addressing
DCS4	Ability to manage the tradeoffs between data collection and storage requirements	E_DCS4	Whether or not students are able to balance data collection requirements with storage requirements
DVTI1	Ability to create a visualization for a dataset.	E_DVTI1	The accuracy of a generated visualization
DVTI2	Ability to identify which data should be used to address a certain question. This includes identifying outliers and creating rules for the computer to filter outliers	E_DVTI2	The degree to which students' manipulation of data to is appropriate for addressing a question
DVTI3	Ability to use the data to create an appropriate model that demonstrates relationships within the data	E_DVTI3	The degree to which students' models and/or description of models reflects the data and relationships of interest
DVTI4	Ability to interpret data models and visualizations for making predictions or drawing conclusions	E_DVTI4	The appropriateness of the prediction and/or conclusion students draw from a model and/or visualization

Table 3. Possible evidence of the assessment targets listed in Table 2

PHASE 3: DEVELOP ASSESSMENT GUIDELINES

Using the assessment targets and evidence statements, we developed a set of assessment guidelines. These guidelines specify the requirements of any assessment tasks developed and specify ways in which assessment tasks can vary. They are used to ensure that the assessment provides the environment needed to collect evidence that reflects on the desired abilities. For example, one guideline is that students must be given the purpose of the data they are collecting or using. Having this requirement ensures that students will be given an opportunity

to use data similar to a real-world situation. Another guideline is that assessments can vary the number of data types students can choose from when collecting data. This varies the complexity of the tasks that students are engaging in.

While assessment guidelines can be developed for individual assessment targets, we developed a list at the sub-topic level. The overlap between the assessment targets within each category made it so the assessment guidelines could be applied to all targets within a category. The list of assessment guidelines we developed is shown in Table 4

Data and Analysis Strand	Assessment Guidelines	
Collection and Storage	AG1	Students must be given the purpose behind collecting and/or storing data
	AG2	Students must be given opportunities to choose aspects of data collection (e.g., what variables, frequency of data collection, the format of storing the variables)
	AG3	Assessments can vary the number of and type of choices students have for data collection and storage to vary the task complexity
	AG4	Scores should be based on the appropriateness of student choices based on the purpose of the data collection
Visualization and Transformation and Inferences and Modeling	AG5	Students should be provided with a purpose and/or question to address and then either provided with or asked to generate a dataset or data visualization(s) to address the purpose or question
	AG6	Students must be given opportunities to develop a model (or representation) or use a model (or representation)
	AG7	Assessments can vary the complexity of the data, the visualization used, and the relationships in the data
	AG8	Scores should be related to the appropriateness of the data representation used or generated and/or the appropriateness of the inference made based on the data

Table 4. Assessment guidelines for each category of assessment targets

PHASE 4: USING THESE MODELS TO DEVELOP GAME FEATURES

Next, the team narrowed down the list of assessment targets to focus the game on a smaller set. The decision on which assessment targets to focus on was determined by the content experts’ knowledge of the concepts critical to the field and the game designers’ knowledge of what features were feasible to be developed. This is important because the assessment needs must be balanced with the development costs and any game play constraints. It can be challenging to meet all of the guidelines while still maintaining the flow of the game. While it would be nice to include all aspects in a game, there are limitations of time and resources that can be used in the development process.

We identified four assessment targets through discussions on required evidence and how the game could support gathering that evidence while still being an environment where students freely explore and play flexibly (as discussed in Chapter 2) (Table 5).

Data and Analysis Strand	Assessment Targets	
Collection and Storage	DCS2	Ability to identify how to automate data collection (e.g., how often to collect the data, and the use of a computational tool to collect the data).
	DCS4	Ability to manage the trade-offs between data collection and storage requirements.
Visualization and Transformation and Inference & Models	DVTI2	Ability to identify which data should be used to address a certain question. This includes identifying outliers and creating rules for the computer to filter outliers.
	DVTI4	Ability to interpret data models and visualizations for making predictions and drawing conclusions.

Table 5. Assessment targets in Beats Empire

From the original list of assessment targets (Table 2), we focused on two assessment targets concerning Visualization and Transformation and Inference and Models. We knew we wanted

to measure students' abilities to identify which data should be used to address a certain question (**DVTI2**) and interpret visualizations to make a prediction or draw a conclusion (**DVTI4**). To measure these targets, we needed to allow students to choose which data representation they used, whether or not they used data at all, as well as their motivation for correctly interpreting the data representation. When exploring data representations for the game, we wanted to ensure that the game could be used in classrooms with a variety of curricular contexts and/or materials, and therefore did not want to include representations specific to any given discipline. This led us to exclude representations such as databases and queries only used in specific courses. Using the guideline that representations could be either provided to students or generated by students (**AG5**), we decided to present students with multiple data representations of song popularity such as line graphs, bar graphs, and heat maps, and focus on students' selection of appropriate representations. These graphs were chosen as they should be familiar to students in this grade level which means that we would expect students to be able to use the representations. The choice of graph students use should hopefully reflect how students were engaging with the data and not their familiarity with specific representations. Students were given several opportunities to make decisions with data, including making decisions about which artist to hire and which aspects of songs they should choose when recording songs, and making predictions related to aspects of songs they picked. To encourage students to engage with the data, students were provided with opportunities for extra awards if they correctly made predictions from the data representation screen.

For Collection and Storage, we decided to emphasize two assessment targets, which measured students' ability to identify how often to collect the data (**DCS2**) and their ability to manage trade-offs between data collection and storage requirements

(DCS4). We felt these targets were more common across different disciplines and could make the game usable in various classrooms. This emphasis on more flexible assessment targets cause us to de-emphasize measuring how well students were able to pick a representations for storing data (DCS3) since it deals with various storage formats like digital, analog, ASCII, and UNICODE that might not be suitable for the instructional targets of all classrooms. We also provided students with the possible variables they could collect, which meant that students were not choosing which variables to collect data on and instead focusing on the frequency of collection for the given variables. Applying the fact that we need to provide a purpose for data collection (AG1) and that there should be opportunities to make decisions about aspects of data collection (AG2), we designed the game to initially provide students with a limited amount of song data, and later provide opportunities to collect/purchase additional data. Students were given opportunities to decide what additional data they wanted to collect to move forward, and how often they want to collect the data. Students were given a limited amount of money and storage to encourage them to make thoughtful decisions about the data they are collecting and how that data relates to the amount of storage they have.

While Table 3 represents the student model, further work was developed to represent the evidence model and task model. The task model, or what would be asked of the students, is really the premise of the game. We knew we wanted a game in which students would use data to make decisions about managing a music studio. We then needed to determine what actions students might do in the game that would allow us to measure the assessment targets.

Notice that the potential student actions and interpretation of student actions are related to the evidence listed in Table 6. For example, for measuring students ability to interpret data models

and visualizations for making predictions or conclusions (**DVTI4**), our desired evidence as specified in Table 3 above is “the appropriateness of the prediction and/or conclusion students draw from the model and/or visualization.” The potential student actions then are related to how often the prediction and/or conclusion is appropriate based on the visualization that the student was examining.

For measuring the ability to identify the appropriate data to use (**DVTI2**), the evidence we originally wanted (Table 3) was the degree to which the students’ manipulation of the data is appropriate to the question. While students did not directly manipulate the data in the game, we could determine if students were using data and if the data they were using was appropriate for the conclusions they were making.

We were also unable to get as strong evidence as we had originally identified for the Collection and Storage assessment targets. This was due to the game providing few opportunities for students to directly manipulate the variables and data being collected. However, we were still able to get evidence related to the original desired evidence. The student actions did provide evidence of whether or not students were considering the data they were collecting and the trade-offs between collecting and storing that data.

Assessment Targets	Potential student actions	Interpretation of student actions
DCS2. Ability to identify how to automate data collection (e.g., how often to collect the data, and the use of a computational tool to collect the data)	Student going to the collection screen	Provides evidence that the student was considering the data they were using/collecting
	Student making collection action (e.g., modifying the way the data was collected)? If so, what action(s) did they do?	Provides evidence that the student was considering the data they were using/collecting
DCS4. Ability to manage the trade-offs between data collection and storage requirements	Did the student follow up actions where they increased storage by buying more data and/or did they increase collection for one variable while decreasing collection for another variable	Provides evidence that students were considering the relationship between the storage requirements and the data they want to collect
DVTI2. Ability to identify which data should be used to address a certain question. This includes identifying outliers and creating rules for the computer to filter outliers.	How often the student looked at the data visualizations and/or the trend screens before picking song traits	Provides evidence that students were considering the data when making decisions
	How often the student made decision consistent with the data they were viewing (e.g., student is examining the mood screen, and then makes a decision related to the mood of the song)	Provides evidence that the student is able to identify which data would provide insights to which decisions
DVTI4. Ability to interpret data models and visualizations for making predictions or drawing conclusions.	How often the student made a prediction in the game that matched the data they were currently viewing	Provides evidence of student's ability to interpret visualizations for making predictions
	How often the student made decision consistent with an appropriate interpretation of the data they were viewing (e.g., picked a trait for the song that was either trending up or currently the highest)	Provides evidence that the student is able interpret the data models to make a conclusion about the data they are seeing.

Table 6. List of assessment targets (student model), student actions that would reflect on those targets, and how those actions could be interpreted (evidence model)

While Beats Empire does not measure all aspects related to data and analysis, using the ECD process helped us identify which aspects to include and how we could use the game to gather evidence of students' abilities. The assessment targets developed were a starting point for conversations with game developers

about what we wanted to ensure was included in the game and how the game could provide evidence of those assessment targets. Often the discussion was around the trade-offs of limiting students to a particular set of choices in order to ensure that the assessment data was collected on those choices and enabling students to decide which choices they wanted to address. Because we did not require students to play the game in a specific way, we had to create a game that would motivate students to engage with the concepts we were interested in. Finding a balance that allowed us to obtain evidence of students' abilities while still making the game engaging even if players were not able to demonstrate high ability was a challenge. For example, one discussion was around whether or not students would be required to identify the data they used to make decisions. From an assessment point of view, requiring this would ensure that students would have the opportunity to interpret data and they would be encouraged to do so. However this would interrupt the game flow for students who prioritized other criteria for making in-game decisions. Ultimately we decided not to include this type of requirement. We wanted to ensure that students had different pathways through the game and could make different kinds of decisions (instead of forcing students to engage in only one particular way). While this does mean there is not enough information to conclude that these students are able to interpret data, it does provide information on whether students would naturally think about using data. As we will discuss more in chapters 5 and 6, this does mean that caution should be taken when determining how students' decisions in the game relate to their competence.

Another decision we made was around how much students would be allowed to fail. This discussion came up when exploring options for measuring skills related to Collection and Storage and the dependency between allowing students to pick what data to collect and then providing representations of that

data. For example, if students were only given the option of examining data that they selected to collect, we were concerned that the students who did not collect data might get discouraged when they had no data to view. The compromise we made was to start with an initial set of data that was collected, and then to allow students to modify the frequency of the data collection, with the goal that increased frequency would provide additional insights into the trends of the data. This way students are able to start with some data and then make further decisions about the type and frequency of data they are collecting.

The assessment targets also provided a start for the conversation of how the gameplay data would be analyzed. This provided input for what data should be collected during gameplay as well as how that data would be analyzed. Further discussion of this analysis is provided in the next chapter.

CONCLUSION

When building an assessment game, it is critical to clarify the goals of the game – what is to be measured and how to meet these goals; this clarification in turn ensures that the game fulfills the purpose for which it was designed. In the ECD approach developers should:

1. Define assessment targets
2. Specify the type of evidence needed to measure those targets
3. Develop guidelines for the assessment
4. Use these guidelines to define the features of the game.

Using an ECD approach and building the first iteration of assessment targets before the game is designed helps ensure that these assessment targets are integrated into the game. Furthermore, the ECD process encourages discussion between

assessment designers and game designers, providing a way to document decisions that are made throughout the design process. This discussion and documentation ensure clear information on what the game will measure, how this measurement will take place, and what evidence will be collected about the students' performance. Finally, while trade-offs are expected between assessment design and game design, documenting how the game design meets the assessment needs provides guidance for the game's use, analysis of game data, and conclusions drawn from students' gameplay.

CHAPTER 5.

LEARNING ANALYTICS IN AN ASSESSMENT GAME

SUMMARY

One of the critical decisions that must be made in the context of a game for assessment is how to analyze and interpret gameplay data. In our game, all student actions were recorded, allowing us to observe which screens students examined, what data they accessed, and how that data was represented in the decisions they made. This chapter discusses how data was analyzed and interpreted to address questions about how students/players engage with the game and use data within the game.

INTRODUCTION

As a formative assessment game, Beats Empire must provide information to teachers that enables decisions about how to adjust teaching and learning going forward. As mentioned in previous chapters, the Beats Empire game was designed to provide teachers with actionable feedback that reflects on student competencies. In Chapter 4 we discussed initial ideas on how we would be able to draw conclusions about students based

on their gameplay. Once the game was built the assumptions that we used to develop the game need to be further tested. Specifically, we need to determine if the way that we believed students would play the game was actually the way they played the game, and if the conclusions we drew from students' gameplay were accurate reflections of students' abilities.

We tested our assumptions through two activities, one is the analysis of log data which will be discussed in this chapter, and the other is through cognitive think-aloud activities with students which is discussed further in Chapter 6. The analyses for both of these activities are used to address the following questions:

1. How can we categorize different types of Beats Empire players? What are the different ways that students strategize and play Beats Empire?
2. How do students reason with and use data in the game? How consistent is students' use of data?

In the development of any type of assessment there is a stage where the assessment is validated. Typically this involves having students and experts examine the assessment to ensure that the language is understandable, and then having a set of students pilot the assessment to collect data on how they are answering questions. We often look for areas in which students do not seem to be understanding the questions, or where students answer questions using skills that are different from those intended. On an assessment we can often do this type of analysis question by question. However, in a game the variety of ways that students can interact with the assessment make it more challenging to isolate students actions. Students are not presented with a single question at a time, but instead are able to explore the entire game on their own terms. Therefore, a critical component to the analysis is understanding the different ways students interact

with the game and using that data to then frame how students interact with the constructs being measured.

Using Log Data to Identify Unexpected Gameplay Patterns

A key affordance of data mining techniques for learning and assessment games is both to provide evidence that players are experiencing the game as the designer intended, and to reveal unexpected gameplay patterns. These unexpected patterns can reveal issues that need to be resolved in the game design, or potentially productive experiences that the designer simply isn't aware of or looking for.

For example, in an early analysis of Beats Empire log data from a pilot with two middle school classrooms, we found a diverse range of strategies for gameplay (Pellicone, et al, 2019). By plotting the number of actions each player took in the game, the number of times players encountered the “insight” screen (the screen where listener interest data is plotted), the amount of money accumulated over the course of a gameplay session, and the number of times players won or lost the game, we categorized three primary types of players. The first kind of player was described as a low engagement player. These players moved through the game somewhat slowly, making few actions and never playing the game to completion (in either a win or loss). This data does not tell us whether these players are struggling to understand how to play the game or are simply taking more time between each action, but it does suggest there is a group of players that are not engaging deeply with the game. On the other hand, another group of players had a very high frequency of actions. These players recorded a lot of songs, made a lot of money, and played the game more than one time, but they didn't visit data screens very often. We might say these high engagement players played the game as a game, recognizing that more actions would lead to more chances for in-game success. However, they may not have found the insight screen, or believed

they could be successful in the game without drawing inferences from this data. Finally, a third category of player also had a medium to high frequency of in-game actions, but they also engaged much more frequently with data representations found on the insight screen. These players may have also played the game more than once, but they tended to reach the victory screen more quickly than their counterparts.

As designers, we certainly intended players to interact with data representations throughout their gameplay. Game mechanics are optimized when players reflect on game data and win conditions are more easy to achieve for those that successfully reason with data representations. And yet, it's also a game! And so we should expect players may play the game in unintended, or at least, unexpected, ways. Consequently questioning "who played the game correctly" is both philosophically counter to our design of *Beats Empire*, a constructionist game for data analysis, and obfuscates forms of play that game designers and teachers, who are using the game as a formative assessment tool, may find useful for evaluating player understanding. For example a teacher may want to check to see if low-engagement players are struggling with understanding a game mechanic, or may want to informally question a high-engagement player about their use of data in their gameplay strategy to bring this important aspect of the game to their attention. This top-down log analysis lets us see these different players, categorize them, and respond with appropriate support and feedback.

While the previous example highlights a top-down approach to analyzing log data, where the analysis assumes a relationship between logged variables and possibly useful outcomes, we also looked at the data from the two pilot studies using a bottom-up approach to identify possible surprises (Zheng, et al, 2020). For example, we might examine the distribution of song genres students choose to record, determine when in the game (what turn) players choose to release songs, or document how often,

and when, students switch between different graph types (line, bar, and choropleth) in the insight screen. In our bottom-up analysis one bit of data stood out: some players pressed the “Click to generate song title” button a lot. When recording a new song, players click this button to auto generate song titles. These titles are randomly created from a table of nouns, verbs, and adjectives creating song titles as diverse as “His Ambitious Dance Club,” “If Another Rose Knows,” and “One Funky Champion.” It’s a lot of fun to press this button and find out what new song title might be revealed and to reflect on whether or not this bizarre title fits the chosen genre, mood, and topic of the to-be-recorded song! And yet, we know that time spent cycling through song titles is time not spent engaging with data.

One response may be to do away with this feature altogether. Perhaps the title is autogenerated and not changeable, or a list of five titles is offered and the player can choose from one of these. This design would ensure players don’t spend too much time searching for that perfect song title. However, it’s also possible that choosing the song title is a personally compelling activity for some players. In this case, removing the ability to explore song titles would decrease a player’s enjoyment and engagement with the game all together, and with it, their exploration of data. In any case, this bottom-up analysis of the log data makes this player action visible to the designer so that followup analysis methods can be employed—such as a think-aloud or an interview with players—to uncover the likely causes of play patterns and possible implications of game design changes.

Students’ Use of Data in the Game

While the previous analysis focused on how we can categorize different Beats Players gameplay, further analysis focused on categorization of students’ use of data in the game. As discussed in Chapter 4, we developed a list of potential actions that would relate to each of the assessment targets we wanted to measure

(viz. Chapter 4, Table 6.) These actions represent our hypothesis for how students would use data in the game. For example, we hypothesized that students who were able to use data to make predictions (assessment target DVTI4) would first view the appropriate data visualization, then use the prediction mechanisms ensuring that their prediction matched the data they were viewing. Recall that when recording a new song, players can make a prediction about whether a song's genre, mood, or topic is "most popular" or "trending up" (increasing in popularity). If a student made a prediction that "angry" songs were becoming more popular over time (referred to as "trending up" in the game), but only viewed a bar graph of the data and never looked at the line graph we would assume that the student did not understand which data representation corresponded to determining an increase in popularity. If the student did view the line graph, we are still limited in what we can interpret from this action. If "angry" songs were in fact trending down, we might assume that while the student was able to pick the right data representation to look at, they were not able to correctly interpret that data. If a student correctly interpreted the line graph to pick a song mood that was trending up and used that mood for their prediction, then we would assume they both know how to pick the appropriate data representation and interpret that data.

For this analysis we wanted to see if students engaged with the data in the way we had hypothesized, allowing us to see if we could collect evidence that students are able to engage with the desired assessment targets. Importantly, while this analysis provides evidence that players are engaging with assessment targets in a variety of ways, lack of evidence does not mean lack of ability. For example, if a student decided to not make a prediction when recording a song, it was not clear if this was because they didn't know they could make predictions, they didn't know how to make a prediction, or they decided not to

make a prediction for an unrelated reason (e.g., maybe they wanted to release songs quicker and didn't want to take the time to make a prediction). Likewise, students may make a correct prediction on one turn, and then fail to predict correctly on another. To overcome this uncertainty we needed many samples of students' use of the prediction mechanic and we needed to triangulate this data with other metrics and observations.

DEVELOPMENT OF CATEGORIES FOR EACH ASSESSMENT TARGET

To start, for each assessment target, we came up with classifications for student actions. These classifications would allow us to provide feedback to teachers on where students are with relation to the assessment targets. We started by using the in-game actions hypothesized to be relevant to each assessment target and then exploring the frequency and sequence by which students performed these actions. This allowed us to see patterns in how students interacted with the game. From there we used a two step process, with the first step being re-evaluating our original assessment targets to determine if changes should be made to better reflect student actions, and the second being to generate the categories by which we would classify students.

Reflections on the Assessment Targets

From the gameplay we determined that some of our original assessment targets could be combined to generate new assessment targets. This was based on our observations that actions in the game could reflect on multiple assessment targets. For example when it came to collection and storage, we noticed that the actions related to our collection and storage targets overlapped. Therefore we decided to merge these two assessment targets into one categorization of students. Instead of separate assessment targets for collection and storage we had one target: "Ability to identify and manipulate variables that should be collected for a specific situation"

Similarly we noticed overlap in students actions for data visualization, transformation and inference assessment targets. However, we also noticed that students had clear separation between actions related to making *decisions* on what song to record, and actions used to make *predictions* (using the prediction mechanism built into the game). This caused us to re-think how we split up our original targets and instead of splitting the data by if they knew what data to use and if they could use the data we split the categories into the use of data for making decisions and the use of data for making predictions.

CATEGORIZING STUDENTS RELATED TO THE ASSESSMENT TARGETS

Using Data for Collection and Storage

When determining how to categorize players we focused on the observed actions to determine the different ways players demonstrated abilities related to the new targets. For example, we found that students did not visit the collection and storage screen very often, and if they did visit it they only visited it a few times across all of their sessions. Therefore we did not find any meaningful differences in the number of times students visited the screen, so this factor did not go into our categorization. Instead, we focused on the actions that they did when they visited the collection and storage screen. We found that students fell into one of these categories:

1. Did not visit the collection and storage screen
2. Visited the collection and storage screen and did nothing
3. Visited the collection and storage screen, toggled the actions but ended up not changing anything
4. Visited the collection and storage screen, bought more storage but did not change anything
5. Visited the collection and storage screen, bought more

storage and increased the amount of data they were collecting

One note is that the game did not provide a wide range of activities for students to engage with related to collection and storage. We felt that while the game could draw conclusions of students' awareness of data and collection, there were not enough opportunities in the game to be able to draw conclusions about how deeply students were able to engage with the collection and storage screen. Therefore, the categories we developed focus on students' awareness. The categories related to the collection and storage target, along with the gameplay actions are as follows:

Collection and Storage assessment target: Ability to identify and manipulate variables that should be collected for a specific situation

- Category: Not enough information
 - Student did not visit the collection and storage screen or visited it but did not do any actions on the page
- Category: Awareness of collecting data
 - Student visited the collection and storage screen and toggles the collection variable
- Category: Awareness of the relationship between storage and collection
 - Student buys storage and then increases the data they are collecting
- Category: Ability to identify variables that should be collected
 - Student increases the frequency of a variable to be collected and then continues

to use that variable when making predictions

Using Data to Make Decisions

To determine the range of how students used data in the game to make decisions, we generated a graph that coded players actions in the game for each song recorded. As shown in the graph below (Figure 1) a song was coded grey if the player did not look at any data representation, blue if they looked at data but the data representation viewed was unable to give them information that led to the choices made when recording the song (i.e. they viewed a bar graph but predicted a song is trending up), light green if the student chose a characteristic that was most popular based on either the line or bar graph representation they viewed, and dark green if they picked a song that was trending up after looking at the line graph. The red represents when the student lost the game (ran out of money), the yellow when they won the game (at which point they started playing again), and the black represents when students stopped playing the game for at least two hours. Note that for students who were categorized as blue for song we gave them a score of 0 to represent they did not make decisions consistent with the data, and for students who were in the green we gave them a score of 1 to indicate that they made decisions consistent with the data.

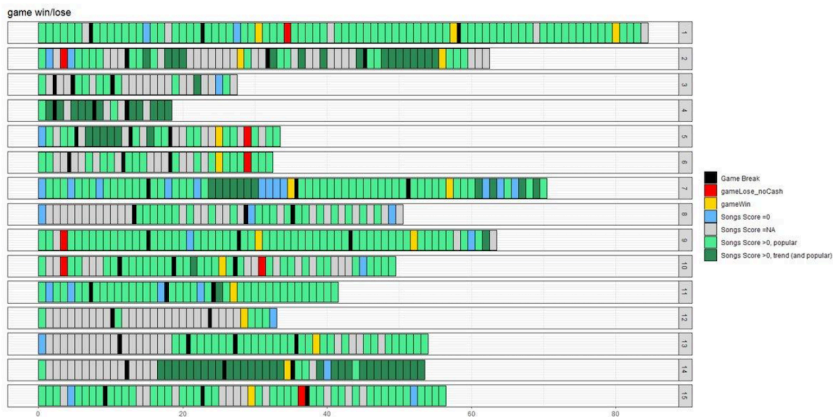


Figure 1. Students' data use when recording songs

Using Figure 1 we are able to look at these sequences of actions to categorize players. We found that there were players who rarely used data, players who mainly used the bar graphs, players who mainly used the line graphs, and players who switched. In our pilots all players made some data-informed decisions about which songs to record, though occasionally players made decisions that were not supported by the data. Being able to categorize a player's in-game actions is useful, but even in an assessment game, players often learn (Chapter 3). Players may take a little while to become familiar with the game and how it works or may gain new insight about how to read data representations after experimenting in the game. Consequently, it was also important to use representations like Figure 1 to document how students' behaviors in the game changed over time. Such representations may also be useful for teachers who want to not only assess the level of understanding of data students have while playing the game, but also how their students are improving. However, providing these representations in a form that teachers can easily understand and quickly act on is not straightforward. More information on how we achieved this is discussed in Chapter 7. In addition to the visualization of student's gameplay data we also provided an overall category of

the student based on the highest level of activity we saw. The categories and corresponding player actions we generated are:

Visualization and Inferences: Students are able to use data to make *decisions*

- Category: Not enough information
 - Student never looked at data before making decisions about song recordings
- Category: Students do not use data appropriately
 - Students looked at data, but the song choices they made were not the highest or one that was trending up.
- Category: Students are able to use a bar graph to make decisions
 - Student focuses on the bar graph (and did not use the line graph), and were able to choose song characteristics that are the highest for some of their songs
- Category: Students are able to use the line graphs to make decisions
 - Student sometimes (or always) uses the line graph and are able to choose song characteristics that are either the highest or trending up

Using Data to Make Predictions

For the last set of categories, we looked at the log data that indicated how players used the prediction mechanisms to make predictions about the popularity, or trends, for the various qualities of songs they were about to record. Recall that many students looked at various data representations when making in-game decisions, but not all players used the prediction mechanic.

Likewise, the prediction mechanic was only available to students who were examining the data when making choices about songs. These students could predict if a song characteristic was going to be the most popular song choice (which they could determine from either the bar or the line graph) or if the song characteristic is trending up (which can only be observed from the line graph). While there is overlap between what we learn about students from the prediction mechanism and what we learn about students from the overall decision they made, the prediction mechanism gives us an opportunity to distinguish if students are making a choice based on if they think a song is trending up, or if it is most popular. In addition, by providing teachers with students categorization based on how they made a decision separately from how they made predictions we gave teachers and opportunity to see if students are consistent across these two different mechanisms. Due to this we kept the same categories (changing decisions to prediction) and based the categories of students on their highest level displayed.

Going back to research question two, from this analysis we found that students had a variety of ways in which they interacted with data in the game. While more students did tend to use the bar graph than the line graph (Figure 1) we do see that some students switched between them and there were students who did use the line graph. We also saw that when making predictions some students used representations consistent with their prediction and some students did not. Unsurprisingly, we also found that students are not always consistent in their use of data in the game.

CONCLUSION

While we developed the game with interpretation of actions in mind, it was important to check to see if how we thought actions should be interpreted matched with how students engaged with the game.

One aspect to note is that throughout this process we were aware that players might make in-game actions for a variety of reasons. Therefore, rather than see the absence of a set of actions as evidence a student does not understand or cannot perform a particular skill, we looked for in-game evidence that players could engage with the skill(s) indicated in the assessment target. Additionally, if we saw mixed evidence (e.g., sometimes students' decisions matched the data they were seeing and sometimes it did not) we chose to interpret this as evidence the student was able to engage with the skill. Further discussion of this is included in the next chapter which also examines student discussion during gameplay to provide evidence for why they were performing certain actions.

The design of the game started with the identification of assessment targets. However, over time these targets evolved, in part due to design constraints (e.g., not being able to include all possible features in the game), and in part after observing how students interacted with the game. In particular, we noticed that some assessment targets could be combined into one target with different student categories associated with this target. For this project we had the luxury of modifying the assessment targets as we were not tied to a curriculum that specified the constructs students were learning.

While we were not able to measure all of the assessment targets, the game provided a context that could be expanded on to gather more information about students. To address additional assessment targets, we developed additional activities that could be administered outside of the game that expanded on the constructs measured in the game and covered additional topics. These activities are further discussed in Chapters 8 and 9.

One of the benefits of working with a game is the ability to collect a large amount of data around players. The challenge then becomes how to analyze and interpret this data. This starts with

identifying which data is relevant to your needs and focusing the analysis on that data. Identification of relevant data can come from a theory of how students' actions reflect on the goals of the game as well as exploration of the data itself to help clarify how students are interacting with the game. Using both of these approaches helps ensure that the game is accurately reflecting students capabilities. This information can then support teachers in their understanding of what their students know and are able to do. Overall, if the game does not provide useful feedback to teachers then it is limited in how it should be used in the classroom.

CHAPTER 6.

VALIDATING THE ASSESSMENT USING COGNITIVE INTERVIEWS

SUMMARY

Beats Empire is situated in an engaging context for students to explore and develop data related computational thinking skills. We conducted cognitive interviews to examine how players engaged with the different types of data in the game, and how players used data to make decisions. In this chapter we will describe students' gameplay, what data students found meaningful in the game, and their decision making process. We discuss the implications for what we learned about these players and what we can say about their knowledge and skills in Data Visualization & Transformation and Inference & Models assessment targets in Beats Empire.

INTRODUCTION

As mentioned in Chapter 5, to ensure that the information provided to teachers is meaningful, we need to determine if what we are saying about students accurately reflects their abilities. This process of collecting evidence to support the validity of

an assessment for its stated purpose is important in all types of assessment and is critical in ensuring that a game-based assessment is able to provide accurate information to teachers. While the learning analytics data, as described in Chapter 5, provided information about what actions students were doing, it does not explain why they were doing these actions. This chapter focuses on cognitive interviews, where students talk aloud as they play the game, to obtain information on why students are making the decisions they did during their gameplay. This chapter addresses the questions of “What data do players find meaningful to define success in the game?” and “How do players use data to inform their decisions?” We first describe cognitive interviews and then introduce three students who represented the range of players we observed. We describe these students’ gameplay and their decision making process, not just for the song choice but also for choosing artists, then discuss the implications for what we learned about these players’ and what we can say about their knowledge and skills.

BACKGROUND AND RATIONALE

Cognitive Interviews as a Useful Tool

As mentioned above, during a cognitive interview, or think aloud, a participant is encouraged to verbalize their thoughts or feelings as they work through a task (Ericsson & Simon, 1984). This information is generally captured through audio and video recording and then analyzed to determine how a participant makes decisions. In a game-based assessment, while we can make inferences about a player based on their log data we cannot know for certain why they are performing the actions shown in the logs. Cognitive interviews can provide a “peek” inside a players’ head of what they are thinking as they play the game, and help designers and developers improve analytic methods of game log data and increase the validity of inferences made of the assessment target.

For our cognitive interviews, we focused on how students made sense of the data to formulate decisions about what was best for their music studio. We differentiate *game decision-making* from *gameplay* in that we define game decision-making as verbal expressions of making choices in the game. In contrast, we define gameplay as applying knowledge of common game mechanics, and how those mechanics can be either maximized or subverted against the intents of the design (Consalvo, 2009).

The cognitive interview included three different parts. In the first part of the interview, players were asked to complete a background information survey that included questions about how many hours they play electronic games and the names of the games they play. After completing the survey, players were asked to think aloud while playing the game for about 30 minutes and provide reasoning for their decisions throughout the game. After gameplay, players were asked follow-up questions to ascertain their in-game goals and gather feedback about game mechanics.

Selection of Players

Eighteen middle school students participated in our cognitive interviews. We chose Riley, Taylor, and Devon as the focus of our analysis because their diverse approaches to the game illustrated how different players engaged with different types of data in the game. Riley, Taylor, and Devon demonstrate different strategies in using different types of data to inform their decisions while playing the game, making them excellent case studies to highlight the multiple pathways in assessing the Data Visualization & Transformation and Inference & Models assessment targets in Beats Empire.

Overall, Riley tended to be data-driven, in that she used listener interest data for making most of her decisions and was careful about how to use and interpret data. Taylor tended to focus on other forms of data not presented in a graph (e.g., the artist stats or the feedback after a song is released). Devon used his personal

preference more than the other two and varied in what data he used when making decisions.

DECISION MAKING WITH IN-GAME DATA ENCOUNTERS

In this section, we describe key areas in the game where players encounter data and describe how each of our example players used or did not use data in these situations. While the game was designed to encourage students to use data when making decisions about the characteristics of songs to release, the game also had other opportunities for students to use data. Students were presented with data they could use when deciding what artist to hire, and they were presented with feedback on their song releases as quotes from the boroughs, the amount of money earned, and the song ranking.

Artists' Screen

As a reminder, on the artist screen, players are presented with both numeric and non-numeric data they can use to decide which artist to hire. Numeric data presented are the artist's salary per week and the level of songwriting skill represented as a rating, in which 1 is interpreted as low skilled and 5 as highly skilled. A definition of each songwriting skill displays as a pop-up box. Non-numeric data presented are the artist's genre, mood, and topic specializations. The player can upgrade any songwriting skill ability, or song specialization using "Current Cash". Figure 1 shows an example of this data using the artist *Envision Gryphons'* resume.

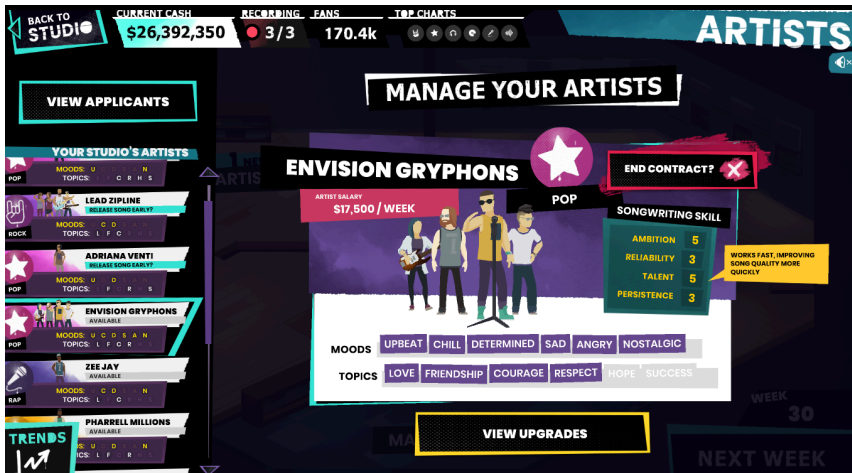


Figure 1. The artist signing screen indicates the songwriting skills of the artist and the moods and topics available for song recording.

Riley. Riley initially stated that when she reviewed an artist’s profile, she looked at the ratings of the talent and reliability to help determine whether to sign an artist. She used her own definitions of these concepts: “I’m trying to decide on talent and reliability because the talent, kind of – you need that...And reliability, it’s like determines, like, how determined they’re going to be to actually do the song well.”

In reviewing the first three artists she signed, Riley chose artists with talent and reliability ratings of 4 or 5. In the case of two artists she chose not to sign, she observed that each artist only had a high rating for one of the skills she deemed important for her artists (talent and reliability).

As she progressed in the game, Riley began to consider how trends in listener interest data interacted with artists’ songwriting skills when making signing decisions. Early in the game, she determined that Turtle Hill listeners were interested in the song moods “nostalgic” and “determined” and in Morris, the trending topics were “courage” and “hope”. Although her signed artists were only able to record nostalgic songs, Riley wanted her

next song to meet both the mood and topic that was trending. In one situation, Riley saw that her two available artists specialized only in writing nostalgic songs, but not in topics about courage. Realizing this, she decided that maybe a new artist could meet these song objectives and reviewed other artists' applications. As she reviewed the artist Brake, she explained to the researcher why she would not sign him, "See like his [Brake's] talent isn't that good. So I would not want to get that. Plus none of the trends follow up with him."

In another situation, when a new artist applicant became available, Riley's review of the songwriting skills and song specialization was met with excitement and she signed him because the new artist was "really good." When asked to explain, Riley says, "Because he had, like, really good talent – it showed that he had good talent. He wasn't that expensive to have – it was, like, only 11,000 or something. And his talent's good, his reliability's good, and his ambition is good. It's just the persistence that I feel like he can work on. But he has a lot of the good traits that he's supposed to have. So that's good."

According to Riley, "really good" meant the cost for the artist was what she interpreted as low, the songwriting skills for talent, reliability, and ambition had ratings of 4, and the artist could write songs about courage. While there was an option to upgrade an artist's songwriting skills and add additional moods or topics, Riley did not take advantage of this option. It is not clear if she did not know this option existed, or instead chose to focus on the skills each artist already had.

Taylor. Taylor shared some similarities with Riley, but also displayed some different priorities and data skills. When determining which artists to sign, Taylor emphasized that he "needs to rely on his singers' talent to try and get some people to like my guys." However, Taylor also considered the other characteristics such as reliability and ambition. For ambition

though, Taylor did not always see higher numbers as better, “So on here, ambition says, ‘Takes risks to create great or terrible songs.’ Right now, I’m not really wanting to take risks. She’s [Madam Zaza] going to take more risks”. Here you can see Taylor bring in some of his own interpretation of how this aspect would affect the songs produced.

Taylor also considered the tradeoff of how many different moods and topics the artist could record to how much money they cost. Taylor expressed his strategy of how he would select artists: “So I need different types of moods in order to get the best, uh, music. And I’m also trying to get to sign artists to the smallest amount of money in the biggest thing.” This statement suggests that Taylor was going to use the “biggest bang for your buck” strategy in signing artists – that is, look for artists with high talent ratings and have a weekly salary that was low. His comments illustrate this, “Right now, for me personally, I’m just looking for the artists with the most amount of talent, because in the beginning talent is important. So his [BPS] talent is 4. His reliability is 2, but his persistence is 2 compared to this one [Manic at the Bistro] whose [persistence] 1. So, I’m going to sign this one...Um, I picked BPS because he had the same exact skill as this one [Manic at the Bistro], except that his persistence was 1 over.”

However, Taylor was unable to hold to this strategy due to the game’s limitations. After the first artist was signed, there were fewer number of artist applications available to choose from. In signing the second artist, Taylor reluctantly signed the second artist – that had low talent (1) – because he only had one artist applicant available to select and he could not advance to the next week until he had another song recording. When it came time to sign a third artist, Taylor was in a difficult position of needing to sign another artist with only two applicants to choose from and a low amount of cash available. Taylor’s reasoning to sign one artist versus the other came down to the artists’ reliability and ambition ratings. “This looks like somebody I need because

he [Reesus Peaces] has a lot of reliability and that's something that I really need right now. I don't think I need this because she [Madam Zaza] has a lot of ambition and that's not what I need right now. I don't really want to take any risks."

Taylor's reservation on signing the artist with high ambition came from his awareness that his studio was in jeopardy of bankruptcy and he wanted to be conservative in spending. After signing the required three artists, Taylor did not sign any more. Instead Taylor focused on upgrading each of his current artists. He first upgraded his artists' talent, reliability, and persistence to a 3. It also became evident that as Taylor's cash began to increase, he was willing to take more "risk" in which he started to increase the ambition skill – a skill he thought was a liability when he had a limited amount of money.

Devon. In comparison to Riley and Taylor, Devon's reasoning for signing many of his artists was ambiguous. Devon started by signing all five artists available to him, without much discussion of the traits of the artist. In signing the sixth artist to his studio, we see that Devon reviews the songwriting skills data for an artist, as he states, "But she [Madam Zaza] looks like she has a lot of ambition. Only thing is the problem is the 1 [for talent]. I'm gonna have to upgrade her, but I'll sign her." When we asked why he signed the artist with very low talent, Devon replied, "I can upgrade her," which suggests that an artist's original set of skills was not a breaking point for Devon, who based his decision on personal preference.

On the other hand, Devon seemed to look at upgrades as an option when he personally felt that it was needed for an artist. For instance, when asked why he was upgrading skills for a one artist, Devon stated the skills "has to be even." He further elaborated why this was important by describing that he felt "there has to be a balance" across the skills. Devon continued using "balancing" upgrades with other artists in his studio. It

should be noted that when Devon balanced the songwriting skills' ratings, he did not upgrade the ratings to the maximum rating (5), but balanced the ratings to the highest rating that already existed (e.g., if the highest rating was a 3, Devon would upgrade the other skills to a 3). This suggests that although Devon understood that low ratings (1 or 2) were not desirable for an artist, Devon did not formulate a relationship that the higher rated artist may produce better songs that generate more revenue.

Devon was the only player who upgraded an artist's song specialization in topic or mood. Devon exhibited a lot of enthusiasm for one particular artist named Brake, so he always had Brake recording songs. After looking at the bar graph on topic trends, he observed that songs about "love" had the highest bar. Even though Devon had an artist who wrote love songs, Devon was determined that Brake would record a love song. Although Brake did not specialize in love songs, Devon decided to upgrade Brake's topic specialization by adding love songs.

The artist screen provided spoof profiles of familiar artists to each player, such as Beats' hip hop artist Brake which is similar to the hip hop artist Drake. The familiarity of the artists provided high engagement opportunities for players to explore data about different artists. Riley, Taylor and Devon demonstrated careful consideration of the data in the songwriting skills, indicating this data was the most meaningful to them to inform their decisions on the artists to sign to their studios. It is interesting to note the different perspectives and interpretations on what skills were important for an artist to have from each player.

In the next section, we will illustrate how Riley, Taylor, and Devon engage and analyze data as they encounter data visual representations.

Recording and Insights (Trends) Screens

When players first go to record a song they see the recording screen (Figure 2), where they can pick their artist and see the current stats of that artist. Players can then click on the Find Trends button to move to the insight screen (Figure 3)

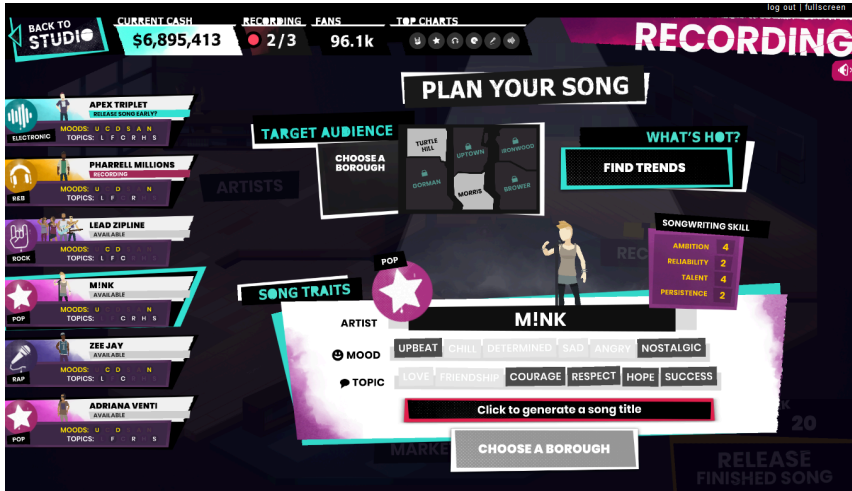


Figure 2. The recording screen where players indicate the characteristics of the song they are about to record.

The insights screen provides representations of the data for the number of listeners by song mood, topic, and genre for each unlocked borough. Each borough has different interests and listener data. For example, when the game starts players can view data for either Turtle Hill or Morris. People in Turtle Hill only care about the mood of the songs, so players can see listeners' preferences of mood but no data is shown for preferences of topics (Figure 3). Morris, on the other hand, only cares about topic and so in Morris, players can view data for topic preference but not mood (Figure 4). Players also have the capability to select the type of data representation (bar graph, line graph, or heatmap) they would like to see. In our pilot studies we found that students rarely used the heatmap (few students clicked on this representation, no students made decisions while using this

representation). While students did click on the line graph, most of their decisions were made while looking at the bar graph.

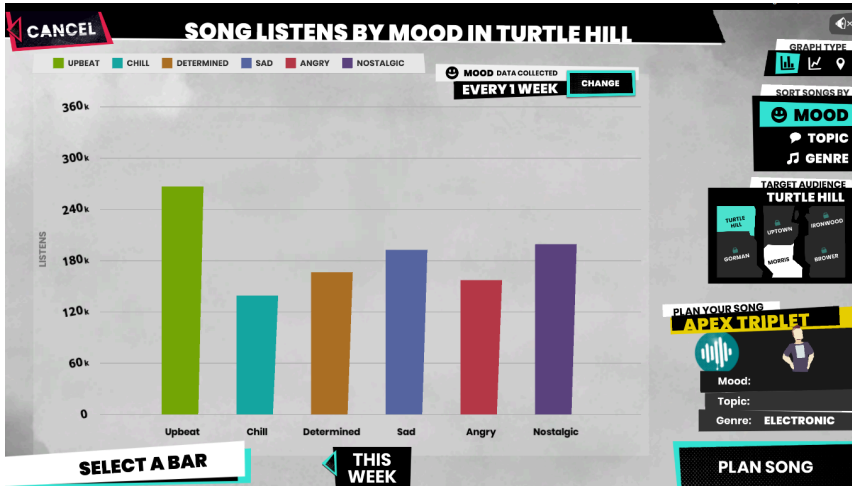


Figure 3. A bar graph of song listens by mood for the Turtle Hill borough.

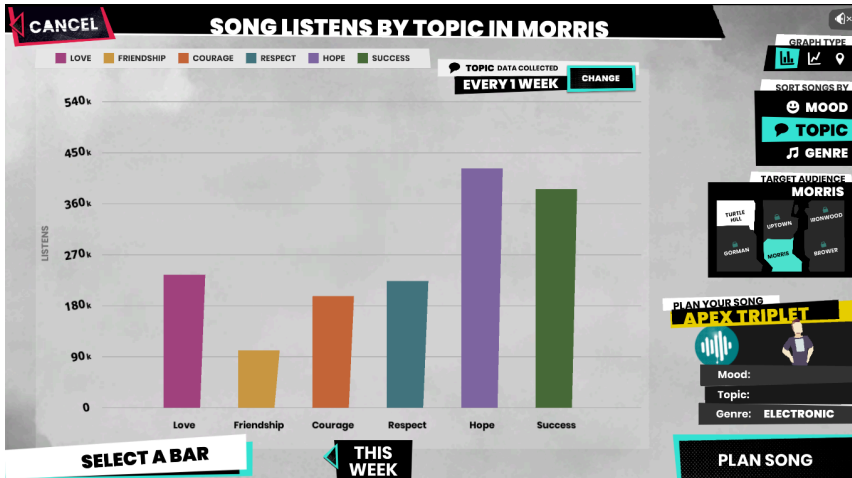


Figure 4. A bar graph of song listens by topic in the Morris borough.

Riley. Riley was very consistent in opening the trends screen to look at data to determine what song characteristics to select before recording a song. During the length of gameplay, Riley recorded 14 songs – seven released in Turtle Hill, five in Morris,

and three in Gorman. While Riley did look at the line graphs once, she quickly went back to the bar graphs.

Riley at times would begin her process to record the next song by saying statements like, “What’s hot?” or “What’s popular now?” Initially, Riley only used Turtle Hill borough data to look for trends to determine songs to record. By the fourth song, however, Riley began to use Morris borough trends in her analysis to determine what song characteristic to choose and which location to release the song. Riley eventually decided to unlock another borough, Gorman, to get more money and more fans. Gorman listeners care about both mood and genre. Using more data, Riley looked at trends across the three different boroughs. Below is an illustration of Riley’s process of tracking and coming to the decision of what song the rap artist Half Dollar would release in Gorman. Note the high cognitive load Riley shows in her tracking the trends across multiple screens.

[*Viewing Trends Screen for Gorman*] And they care about rock and they care about – Oh! They want upbeat now!

[*Goes back to Recording Screen*] How do they make it – oh upbeat
[*Selects upbeat*] and... Do they care about the topic?

[*Viewing Trends Screen for Gorman*] They don’t care, like Turtle Hill. They don’t care about the topic. But what about Morris?

[*Viewing Trends for Morris*] It’s still courage and hope.

[*Viewing Recording for Half Dollar*] So upbeat and courage, ‘cause I’m trying to get it to play at all venues. And since the other two just don’t care, I might as well just do it like that. Let’s do Gorman. [*Selects Gorman.*]

[*Presses Start Recording.*]

We see a shift in Riley’s approach to recording and releasing

songs after unlocking Gorman. Riley looks across all data and sees how to satisfy all the boroughs' trends. First, Riley used the data to determine which artist to record rather than using the artist that just released a song. The second shift was in the signing of new artists to her studio. As mentioned earlier, Riley would only sign artists with high talent ratings. Riley started to relax this constraint, and turned her attention to a new artist's ability to write songs related to moods and topics and into which genre the artist was categorized.

These statements were followed by her looking at the insights screen to see what was trending. It was apparent that when Riley started to formulate relationships between specific song characteristic trends for a specific borough, she got positive results from certain artists. For example, early in the interview Riley saw that the moods nostalgic and determined were trending in Turtle Hill. Two of her artists (Zee Jay and Juno Mercury) had the skill to record songs using one of those traits, so Riley only had those artists release songs in Turtle Hill. Similarly, Butane Plan had the ability to record songs about the trending topic courage in Morris, so Riley would release only their songs in that borough.

Taylor. Taylor recorded seven songs during his interview, which was the fewest number of songs recorded between all three players. Taylor's approach used only bar graph trends data for one borough, Turtle Hill, and the feedback he received on the results screen. Observations seemed to suggest that he had a limited understanding of how to navigate the insights screen. When a song completed its recording to be released, players received feedback from unlocked boroughs such as Turtle Hill and Morris as shown in Figure 5. Taylor seemed to find value in the feedback as he reflected each time it was given. For example, after receiving feedback that his song, "lacks something," on his first released song, he said, "Well, it was obviously not bad for the first song. And I got some new fans, but – I understand why it

didn't get Morris because Morris doesn't really like mood. But I was hoping for something better from Turtle Hill. They said, 'this still lacks something.' So I'm going to have to try and figure out how to get...uh...stuff."

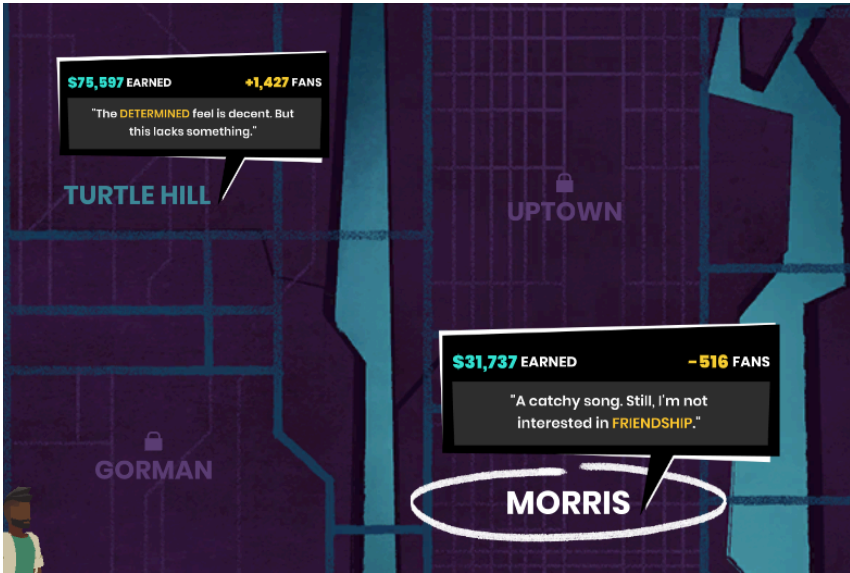


Figure 5. After releasing a song, dialogue bubbles from each borough where the song is released provide feedback to the player.

Taylor used trends data for Turtle Hill to select the mood of the song, however when it came to selecting the topic for the song, he used his own intuition on what to select rather than the trends data. For instance, when asked why he selected releasing a song in Morris, Taylor reasoned, "So I saw that in Morris, the population cares about the topic. And friendship is obviously a topic – that are not obviously – but I think friendship is something that they would enjoy." When asked about the feedback for that song, Taylor seemed to indicate satisfaction in the results: "So I think the results – it's kind of what I was looking for. You see that, um, Turtle Hill only got 1,300 new fans for me and Morris got 10,400. And I was just targeting Morris so that's good. I got some new fans, but it still says, 'It lacks something.' I need to try to figure out what it lacks."

This puzzlement over why he is still receiving that his songs “lack something” in Morris is an illustration of how Taylor was not able to connect how he could have better utilized data in the game to inform his topic choice. Although Taylor was able to use trends data to select a mood for his songs to be released in Turtle Hill, he did not indicate that he should have topic data or any data for Morris. In relation to this, Taylor did not seem to realize that he could have made better decisions in song topics based on trends data for Morris rather than using his own judgment to possibly receive better feedback from the boroughs.

Devon. Although Devon was consistent in using the insights screen to look at data to determine what song characteristics to select before recording a song, Devon did not use all data trends available to him most of the time. Devon’s decisions at times were not seen to maximize the potential outcomes (e.g., selecting the second highest trend mood for a song rather than the highest trending mood), but were more geared towards personal preferences based on his familiarization of the artists and genres. During the length of gameplay, Devon recorded nine songs – six released in Turtle Hill, one in Morris, and two in Gorman.

Closer examination of why Devon focused on recording and releasing songs mostly in Turtle Hill seems to be associated with a couple of motivating factors. First, Devon only used the Turtle Hill data to record and release his first three songs. Another factor was that feedback from the results screen seemed to motivate him to continue to record and release in Turtle Hill. The combination of the comments from Turtle Hill (e.g., “This song’s awesomeness is off the charts!”), placement on the top charts, and the increase in cash were all viewed as positive feedback to Devon. However, Devon did not agree with the comments from the other boroughs. For example, after reading the comment from Gorman, that they “Don’t care for rap.” Devon responded in disbelief, “Who are these people?”

Devon displayed a pattern of using the same artist that just released a song to record the next song. For example, when Zee Jay song was released, Devon would use data to record a new song for Zee Jay. Devon also tended to reuse his favorite artists. For example, Devon decided to record and release a second song for the artist Brake. From the very beginning of gameplay, Devon proclaimed that “Brake is gonna give me a lot of money!” and “Brake is going to be the one who hits the top [of the] charts.” To determine what song Brake would record next, Devon first looked at the mood trends for Turtle Hill and then Gorman. Brake did not have the moods that were trending in both of these boroughs so he decided to look at trends in Morris.

As Devon reviewed the trending topics in Morris, he saw that love had the highest trend, and that Brake did not have this skill. Brake however had the ability to write songs about courage which was also trending up but not as high as love. At this point, Devon decided to upgrade Brake and added “love” to Brake’s resume so he could record and release a song in Morris. What is notable about Devon’s upgrade decision for Brake is that Devon was originally looking at the moods, and he could have stuck to looking at moods and upgrading Brake, but instead switched his strategy to focus on topics and then did the upgrades there.

Devon’s tendency to bounce around on his strategy can be seen in his recording strategy for the next song. At first, Devon decided his next song would be recorded by the pop artist, Mink, and the target audience would be Gorman. Devon noticed Mink’s songwriting skills were not “balanced” (i.e., 4, 1, 5, 2), so he decided to upgrade Mink’s songwriting skills. Once he returned to the recording studio he opened the mood trends for Gorman and saw that Mink had the ability for one of the high trending moods, but found pop was the third highest genre and R&B was the highest trending genre in the borough.

Devon [Viewing Gorman genre trends]: Dang it. We need R&B now. I can't do this with Mink.

Interviewer: Oh okay.

Devon [Viewing Artists screen]: We need some new artists.

Interviewer: Okay, so you need ... oh, so you—

Devon: Pharrell [*last name Millions in the game*], Here we go. [*Views ratings of 3,2,2,1*] I have to upgrade him.

Devon proceeds to upgrade Pharrell Millions to balance the songwriting skills so they all have the rating of 3 then goes back to the recording screen. Devon selects this artist and views trends data for mood and genre in Gorman. Pharrell has the ability to record “upbeat” songs which is the third trending mood and Devon selects that mood and records targeting Gorman. Devon could not select the top two trending moods because the artist did not have the ability to do so, and while he could have upgraded Pharrell to include the highest mood he decided to just focus on the genre.

Decision Making Foci

Riley, Taylor, and Devon were all able to use data in the game to make decisions, however, they used data in different ways. For example, Riley and Taylor invested time in reviewing the artists' songwriting skills because they wanted artists with high ratings in “talent”, while Devon did not critique an artist's resume as much because he knew he had an option to upgrade skills at any time. Riley looked across boroughs and song characteristics to make decisions on which song to record, while Taylor focused on just the mood trends for Turtle Hill and Devon alternated between which type of data he used to make decisions. We categorize these as data driven, talent driven, and personal preference driven and describe them further below.

Riley's data driven focus. Riley's awareness that the game involved analyzing and interpreting data, put her in the mindset that she was going to use data to achieve her goals in the game. As Riley plays the game, she exhibits a high value to use data to inform her decisions such as recording a song or signing a new artist. As a way to regulate herself in the game, Riley seemed to intentionally remind herself that she was looking for trends in the data by incorporating in her language words that related to "trending" such as "popular" or "in-style" as she analyzed and interpreted data. For example, in a few instances on deciding what song she should record next, Riley would say, "What's hot right now?" Riley's engagement with trends and songwriting skills data eventually led her to create songs placing high on the music charts (e.g., platinum and gold).

Taylor's artist talent focus. Taylor's interview exhibited decisions similar to what a music business manager would focus on: having talented artists. Taylor spent a significant amount of time on the artists screen analyzing the songwriting skills data of each artist. Taylor heavily relied on the feedback given from the results and his own intuition to inform his decisions on signing and upgrading artists' songwriting skills. As with running any business, being mindful of money is important. Taylor seemed to exhibit this trait frequently as many statements related to money – spending, cost, losing and gaining money – especially when it came to artists.

Devon's personal preference focus. It was revealed in Devon's interview, that Devon actually had real experiences using data in the music industry as he told the researcher,

"My own mother, she has the same job as in which I'm doing here. So sometimes I go to the studio to help her out...my favorite part is when we go online and look at all the trends like what we do here...So the trends around [a state], but – then – what we're gonna write the song on."

Although the artists in the game are fictional, their names resemble real life artists (e.g., the hip hop artist Brake in Beats is similar to the hip hop artist Drake in real life). Devon projected the same expectations of the real life artist to the fictional artists in Beats Empire. For example, Devon proclaimed that the artist Brake was going to be his money maker. Bringing his own assets and applying them to his decisions, Devon seemed willing to take risks early on in the game and had confidence in his artists' ability to make money for his studio. Although he used data for some decisions, Devon tended to make decisions based on his own preferences without elaborate explanation, and inconsistent with previous statements.

We chose to highlight these three players as they represent a range of styles of gameplay. Other strategies did exist. For example, we found that some students spent a lot of time choosing song titles and less time on other characteristics of the song, while other students did check the line graphs at times when looking at the data. However, the strategies displayed here demonstrated common strategies that different players used, paying attention to different types of data in the game. Some students focused on the graphical representations of data, others on the numeric data shown for the artists, others on the qualitative feedback given after a song was released, and finally other students shifted their data focus throughout the game.

IMPLICATIONS

Drawing Conclusions

The big question after the cognitive interviews became “What can we say about students, knowing that sometimes students change strategies, and sometimes students do not take full advantage of the data they are provided?” We recognize that players are inconsistent with their data use (we all are), and the game supports this by allowing players to engage with data to

varying degrees at different times. The information learned from the cognitive interview supported our previous decision of categorizing students by the behaviors they display, even if this display of behaviors is not consistent. (See Chapter 5). They also supported the assumption that a student could have a skill they did not display through certain behavior. For example, if students chose a mood that was the top mood shown in the bar graph we would indicate that there was evidence that students could interpret bar graphs. We would still say this even if students later picked a mood that was not the top mood represented in a bar graph, as we know that some students will switch their strategies and may not always want to pick the top mood. Similarly, if we do not see students ever picking the top mood we would say that we don't have enough information about the student to draw a conclusion.

The cognitive think-alouds also highlighted the importance of viewing students' behavior over time, and in doing so, indicated that this was information that might be useful to teachers when they are making sense of their students' gameplay. Further discussion of the representations of this data and how we wanted to display them to the teachers is shown in Chapter 7.

Game Design

Data from gameplay logs and the cognitive interviews make it clear that the game allows students to explore freely, and to play using a variety of strategies. Furthermore, by providing frequent game elements where players encounter data, players have many opportunities to make data-based decisions. However, the cognitive interviews did indicate that some improvements could be made to the game.

Providing more incentives to explore the insights screen and the additional representations would improve our ability to observe skills that indicate learning. While heatmaps were included in the game, players did not see a reason to use them. Additionally,

while players occasionally used the line graphs, they often decided to stick with the bar graphs as they believed this representation gave them all the information they needed and was easier to interpret. It may be that players were not fully aware of what options were available on this screen, let alone whether or not those features would be helpful for decision making. For example, Taylor only released songs in Turtle Hill, the default borough, and never clicked on additional boroughs to view data about listeners' interests in these other locations. It is unclear what he would have done if the additional options on the insight screen had been more accessible to him.

It's also clear the feedback screen, displayed after releasing a song, was salient to players. Each player noted the pop-out quotes presented on this screen and incorporated this feedback in their next recording session. However, this screen was not a design priority for the team and the language used in the feedback was not carefully chosen. The cognitive interviews suggest this screen might be an important tool for nudging players towards data-based decision making.

While we can imagine many ways the game could be modified to encourage students to engage more deeply with data and have opportunities to explicitly display evidence of their abilities, the trade-offs of these modifications must be considered. If game modifications over-constrain gameplay towards a data-driven focus, some players may not be able to play the game as they want, which might affect their motivation and engagement with the game. This also brings up the role of the teacher when students are playing the game. While it may be challenging for a teacher to have individual conversations with all of their students, the information provided from the game can point out possible conversations teachers can have with their students to better understand their thinking, and to point out aspects of the game that the students might want to try. More information

about how the game can be expanded to include classroom activities is discussed in Chapter 8.

Overall, the cognitive interviews helped us to better understand the multiple pathways that players used during the game. It was a way to check if the conclusions we would draw about the players were appropriate, and helped us to identify the appropriate level of feedback given to the teachers. While the gameplay data is appropriate for classroom discussion, more information about why students used the strategies they did would be needed before strong conclusions can be made about what students know and are able to do in beyond the game.

CHAPTER 7.

PARTICIPATORY DESIGN FOR DEVELOPING A FORMATIVE ASSESSMENT DASHBOARD

SUMMARY

Beats Empire does not function meaningfully unless the teachers using it understand what gameplay tells them about learner knowledge and understanding, and how this information can inform lesson planning and future teaching. In synthesizing accepted design practices for classroom technologies with deeper investigations into teachers' understanding of and interest in dashboards and data visualization tools, we created a dashboard that is relevant and adaptable for teachers. This work was both enriched and complicated by the intent of Beats Empire – to present computing concepts in a rich authentic context with utility across a variety of classroom domains ranging from math to physical education.

INTRODUCTION

Formative assessments are activities where the students or teachers use data from student performance to change their

learning path (Black & William, 2009). We have discussed how the design of Beats Empire intended to elicit contextualized performance of Data and Analysis understandings from learners in a classroom – in a way that better situates their understanding and thinking than typically found in most classroom assessment activities.

Translating the information from such assessments into actionable insights is a difficult pedagogical task that, like most teacher activities, lies at the intersection of content knowledge, pedagogical experience, and pedagogical content knowledge. This challenging task is further compounded when the activity and assessment are mediated through technologically-driven activities like Beats Empire. Since Beats Empire situates content knowledge in a setting somewhat unfamiliar to teachers, it was our imperative as the game’s designers to develop an initial bridge between the meaning of students’ in-game activities and what it means in the context of teachers’ pedagogical techniques and goals.

We designed and created a dashboard for helping teachers use Beats Empire across a variety of classrooms (herein described as a **cross disciplinary** game and experience). Reaching toward one of Beats Empire’s underlying goals – make data literacy assessment activities accessible and useful in a variety of non-computing classrooms – became a challenging design imperative for this dashboard: to provide information to teachers across different domains about data concepts in ways that make sense to them, and are actionable as well. This involved engaging in extended participatory design with diverse teachers, and engaging in iterative, low-fidelity mockup development and testing. In this chapter, we describe the design principles for learning dashboards that are discussed in prior literature. We then describe our design process and iterations regarding our cross disciplinary dashboard (mirroring the intended design goal of our game), and conclude with key takeaways around designing

such discipline-agnostic teacher dashboards, especially in designing alongside teachers.

Why make a cross disciplinary dashboard? We made a cross disciplinary game for three reasons:

1. Computing concepts like data collection and analysis have cross disciplinary value.
2. Reducing perceived disjointedness between different disciplinary skills and activities enriches and amplifies learning experiences.
3. As computing is a new discipline, schools with differing access to resources and expertise in teaching computing would benefit from having expansive suites of activities which reinforce computing learning in and through other classrooms.

These reasons all reinforce the need for a cross disciplinary dashboard that enables student gameplay to be used as a data source in varied classrooms.

How to make a cross disciplinary dashboard? First, a good dashboard is only as good as the game it is built for. In this book we have discussed extensively what makes a good educational game, but three design characteristics are worth highlighting here:

1. Prioritize addressing educational inequity and work to understand the learning context, audience, and different stakeholders (parents, teachers, etc.) (see Chapter 2).
2. Ensure that the game reflects your learning goals and values (see Chapters 3 & 4).
3. Understand the range of learning goals your game can speak to and how the game encourages students to engage with these learning goals (see Chapters 4 & 5).

In dashboard construction and design it is imperative to:

1. Engage with your intended audience (in our case, teachers) to explore how they understand educational games and dashboards.
2. Identify gaps and connections between the game's ability to elicit behaviors and participation, and the data's robustness for actually speaking to student understanding and interests.
3. Understand ways to leverage connections between known design principles for data presentation and dashboards; and the specific use contexts of relevant teachers monitoring gameplay in the classrooms of interest.
4. Iteratively design and test designs with teachers.

CENTERING TEACHERS AS PARTNERS AND CONSUMERS

Formative design research, conducted before we began designing and building the game, informed us of the different kinds of teachers interested in leveraging formative assessment data regarding computing/data thinking concepts (see Chapter 2). This process was vital as it allowed us to center the needs of our stakeholders throughout the design process and with the final product. Additionally, our work with the teachers allowed us to explore in more depth the ways in which our dashboard could be used as a means to combat persistent inequalities that exist across districts, schools, and classrooms.

We engaged with teachers through an iterative participatory design process. Working with a total of 11 teachers across three separate focus groups, we strove to understand their needs as managers of classrooms, experts in their respective content areas, and experienced facilitators of learning this content

(categorized in the PCK framework – Cochran et al., 1993 – as pedagogical knowledge, content knowledge, and pedagogical content knowledge respectively). The groups were separated by disciplines which we expected to have different amounts of explicit experience with integrating data and analysis into their classrooms. This process is explained in more detail in Chapter 2.

During this workshop, we began to identify trends among the teachers' activity ideas that tied to the learning goals within the CSK12 Framework. Physical education teachers create data-based activities for students to track their food consumption and nutrition information. Language arts teachers ask students to construct arguments drawing inferences from data and theories, which often uses multiple explicit and implicit data literacy skills (Vahey et al., 2012). Math teachers use data inferential activities in many different topics including functions, graphs, and statistics. These use cases and practices overlapped with the numerous learning sub-goals described in blueprints and learning frameworks for different disciplines and topics. For Beats Empire, we centered the K-12 CS Framework as a guidepost for including valuable learning goals in the design of our game, as discussed in Chapters 4 and 5.

By analyzing the teachers' outputs from this workshop, we also developed three personas (Cooper et al., 2007) – a common interaction design method that involves creating virtual personalities to provide compact but abstract coverage of different intended users. Figure 1 presents one of the personas we developed – a Health teacher. These personas highlight and embody different pain points, needs and wants teachers using Beats Empire might have. For example, the Health Science teacher persona is described as having limited experience with computation and needs other teachers to help them integrate computing into their classes. This makes developing effective formative assessments for them difficult as they have limited

experience with the content, and renders a dashboard that can interpret and offer assistant imperative.

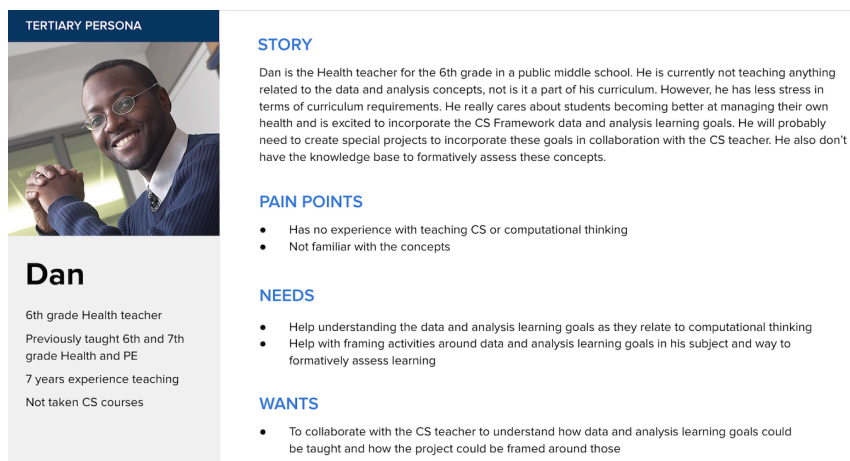


Figure 1. Persona of a Health teacher seeking to integrate data and analysis content into their classroom.

We additionally used information from and about the teachers as a way to combat the persistent inequities that exist in computer science across districts, schools, and classrooms. When creating the dashboard, choosing what is “useful” learning can center extant inequities by valuing sub topics which are not equally accessible – through the contextualization offered in-game or classrooms, or ways that these topics are assessed . The same learners may excel in other undervalued topics, using them powerfully in personally relevant ways. For instance, Turkle & Papert (1990) describe how thinking of object oriented programming as the “best” way to learn and engage in programming harms learners who prefer different ways of thinking and practice, and also harms the discipline of computing itself by being constrained in the problem-solving strategies afforded. With Beats, we intended its real world context and its multidisciplinary uses to be key ways of enabling learning of topics in contexts that are specifically engaging spaces for (computing) minoritized learners, and surface

professional and aspirational opportunities they can more readily identify with, in their learning experiences.

Similarly, in working with and centering teachers that serve minoritized learners, we strove to develop design guidelines for dashboards that discourage further marginalization, and center guidelines to positively serve teaching and learning practices for their students. For instance, the negative effect of comparative grades and scoring practices on marginalized and underperforming learners is recognized across a variety of contexts (Betts & Grogger, 2003). More recently, multiple studies on dashboards with different kinds of comparative scores have also been found to be demotivating for learners (Jivet et al., 2017), especially those in need of additional support. This is a useful reminder to be cognizant of prior learning design practices and their recognized effects, such as avoiding any comparative data among students in dashboards.

The most valuable take away from these participatory design activities was that we identified that significant and careful effort would be needed when designing the dashboard. Initially we had thought of the dashboard as a set of data points from the classroom gameplay, but these participatory activities suggested that the teachers were unsure and even anxious about using formative assessment from gameplay and would therefore need guidance on both content learning gains and engagement from the students. The next steps were to understand how the gameplay data related to learning and research on the functionality and user experience.

UNDERSTANDING GAMEPLAY DATA

The factors explored in this section demonstrate a commonly found design constraint in data presentations for experts (recently frequently studied in AI-equipped diagnostic tools for doctors, as per Wachter et al., 2017), regarding *certainty* of

inference. Gameplay data is unique, especially in the context of learning environments. It often presents a dense stream of decisions similar to cameras, microphones, or other sensor data, but carries contextualized meaning as afforded by the design of the game. In *Beats Empire*, we get information about which screens each player is visiting, how often, and which buttons they are clicking, as well as the state of their music studio and what their decisions may indicate about their use and understanding of their data (described in detail in Chapter 5). Our gameplay data, however, cannot give explicit access to understandings and intentions being exercised from the learner’s point of view. The tension between individual priorities and game knowledge and the game’s metrics is therefore critical and requires sensitivity.

Early research from playtesting revealed how real-world elements in the game surfaced learners’ non-disciplinary interests and feelings about music, leading to decisions with respect to the game design’s success goals and metrics (i.e. earning more money, or making “popular” records) (Pellicone et al., 2019; Thanapornsangst et al., 2020). We know that learning environments and activities that engage learners in the context of their own values and interests – for instance, deciding to sign a pop artist even if they recognize that rap music is more popular in the current state of their game’s world – are a valuable way to engage minoritized learners and help them identify with the discipline’s activities (Ladson-Billings, 1995). Thus, we want to ensure these kinds of decisions are not seen as undesirable, at the game design level or the dashboard-assessment level. At the same time, we also want to ensure that the data translation and inferences our dashboard presents do not misinterpret these kinds of decisions as informed or misinformed if we have a sense of the different reasons learners could have made certain decisions.

Additionally, game and interaction design is not equally transparent and easy to navigate for different players. It is critical

not to conflate the shortcomings of a design's comprehensibility for players'/learners' competence. In *Beats Empire*, for example, we see some players never changing the graph type, or never visiting the storage and collection screen where they can change how much data is recorded about the city's music interests. One might conclude these learners do not understand how different graph types or frequency of data collection is useful for data analysis. Or, perhaps players simply don't notice the UI elements that allow them to switch to different graph types! It is important to identify in what ways our inferences about student competence or preference from data interact with the accessibility of the design of the interface, and student understanding of the domain.

DESIGNING DATA PRESENTATIONS FOR TEACHERS

As an initial step in the design process, we kept in mind the following general interaction design guidelines from Rose (2000), aligned with accessibility and information communication needs:

- Color palettes with adequate contrast
- Fonts to support different size text
- Language suitable to the context
- Text, symbols, and colors used in ways that complement and mutually reinforce
- W3C Accessibility guidelines should be followed as a baseline to make the interface friendly for screen-readers (Galitz, 2007).

Tufte (2016) provides an oft-cited set of guidelines to design effective and productive data visualizations, which include avoiding confusing presentations that increase the likelihood of misinterpretation; maximizing *data-ink* (i.e. the display artifacts used to represent data) on a screen in comparison to non-data

ink, and minimizing redundancy. This minimization can, however, sacrifice accessibility. It is therefore valuable to keep in mind that while some users prefer to interpret graph labels through color, shape, or text, and enabling access to each modality is desirable, this should not overload the display. These provided some starting points for designing individual visualizations, but needed to be supported to respond to the fuller context of teachers in classrooms during gameplay and outside of classrooms during lesson planning, as well as the context of how gameplay data can be presented and interpreted for productive and accurate inferences.

Certainty

As discussed above, in-game decisions can be affected by the broader context of the game and may not always correspond to the constraints of relevance to specific domain-related performance expected in classroom assessments. This challenge was addressed by presenting the data that can be used to make more certain claims. For instance, one of the most common design features of classroom game dashboards information is about in-game activity (reflected as measures of on-task and off-task activity in other educational dashboards – Matcha et al., 2019). Provided that there is an appropriate development of game data streams and servers which can host and provide real-time access to data, teacher dashboards can reasonably identify spans of time when players have not made any in-game actions. Designers can thus designate a window of time as a marker of idleness and alert the teacher. In-game idleness is therefore a common measure intended to assist teachers in classroom management – one of the most commonly requested and expected features of real time classroom dashboards. The ability of games which produce data to help teachers' judgment of on- and off-behavior can reduce teachers' efforts on monitoring engagement and thereby allow them to focus more specifically on student learning. Additionally, in centering non-punitive

classroom practices, the design of dashboards can remind teachers that the goal of markers of off-screen behavior is not to curb off-screen time, but to pay attention to what students are doing when not engaging with the game. A lot of so-called “off-task” behavior in classrooms is actually integral to learning (Cocca et al., 2009). Therefore, creating environments where dashboards help create non-punitive *play* experiences promotes organic motivation and engagement in the games.

Another common design tactic in dashboard designs is to anchor the description of student behaviors relative to each other. While this is similar to many other common assessment practices like grading on a curve, we strongly believe that this fosters an environment of comparison and competition unhealthy to effective learning. Reviews of numerous student-facing learning dashboards have shown that comparative presentations harm student motivations, especially those who are underperforming (Matcha et al., 2019). Similarly, we expect that providing visualizations that compare students in hierarchical manners inherently encourages comparative perceptions of students, pushing students to strive for “more” or “better”, both of which are can be antithetical to enjoyable gameplay – as well as the goal of formative assessments. We thus ensured that our dashboard does not present certainty in data presentations as a relative construct, and strongly advise other designers to be cautious of producing inferences using algorithms that compare data across students without being very conscious of and explicit about the underlying assumptions, mechanics, and goals of such a comparative algorithm.

Considerations, Constraints, and Revisions

An in-depth analysis of other existing dashboards like GoFormative, Naiku, Socrative, and ClassCraft (Atherton, 2018) provided a starting point for features commonly found in such tools. These features include customizing and creating the initial

setup of the dashboard, creating a teacher profile, giving feedback directly to the students, and models of different types of reports usable by teachers. Complementing this with findings from the participatory design sessions and mapping it to the teacher personas, we identified formative design considerations for the dashboard.

Formative Findings

We found most teachers receptive to using games in their classroom, and they expressed that they wanted access to real-time information about students' gameplay experiences – whether they were stuck, headed in the wrong direction, or some other immediate actionable data. At the same time, they were also concerned that a dashboard could overwhelm them with information, making it inconvenient to the point of becoming unusable. Teachers' thoughts on the kinds of information they imagined being useful surfaced into two key categories: 1) information on individual students and 2) overall class performance.

Individual Student Data	Overall Class Performance
<ul style="list-style-type: none">• Name• Duration of gameplay• Time elapsed since the game started• Completed the activity• Students “stuck” and at what point• Students doing “worse” than previously• Individual student performance	<ul style="list-style-type: none">• Number/percentage of how many have completed activity• General class performance• Common mistakes or misconceptions• Performance improvement from previous session• Performance on specific rubric and learning goals• Photo• Indication of who has logged in

Equipped with the ideas generated through these processes and these categories, we developed an initial iteration and low-fidelity mockups to present to teachers to collect further feedback.

Design Iterations

After this formative analysis, we conducted user tests, additional interviews, and card sorting activities to produce iterations and refinements of the dashboard design.

Prototype 1. We started with a low fidelity mockup (Figure 2), presenting a wide range of data to teachers. This mockup deliberately laid out more data presentations than typically usable, to prompt teachers to share their priorities and preferences across the visualizations. This mockup acted as a scaffolded card sorting, where we presented different named, grouped, and partially pre-sorted data sources, soliciting more specific feedback.

We asked teachers to emulate using the mockup and solicited feedback when they made choices or actions that appeared unclear in intention to us, surfacing the interactions and kinds of information that made intuitive sense to them. We found that teachers sought out and valued information about which CS content needed to be revisited in their classes, and about relationships between game activities and educational standards. They also wanted a reduced amount of information on the front page of the dashboard, but still wanted access to deeper information and “raw data” to retrospectively look at and understand the reasoning behind the interface’s suggestions. Having *opportunities to explore the data in detail* became a recurring theme across future prototype iterations as well.

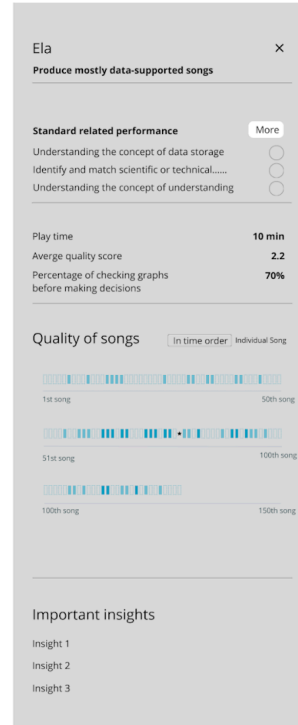
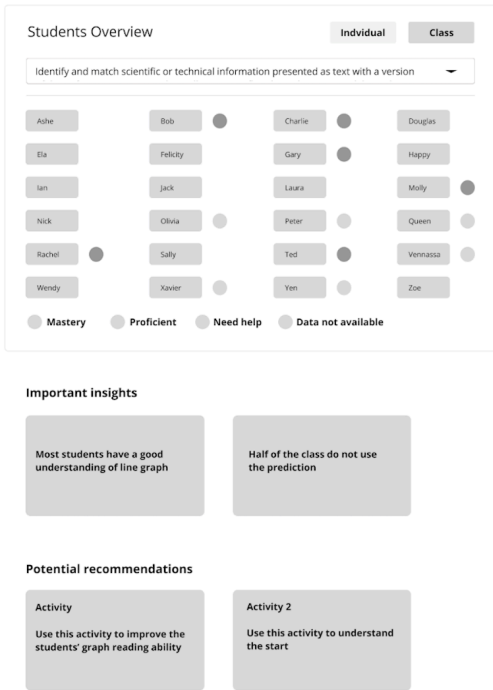


Figure 2. Prototype 1 had a wide range of information included in the dashboard.

Prototype 2. For Prototype 2, we simplified our initial prototype by spreading information across two main windows: *Student Performance* (SP) presenting information about student engagement with the game; and *Important Insights* (II) reporting on student learning. SP information included descriptives and counts of student activity, such as number of turns taken and time on task. We presented a visualization of three of the assessment targets and included ways to interpret and act on the data presented in Figure 3.

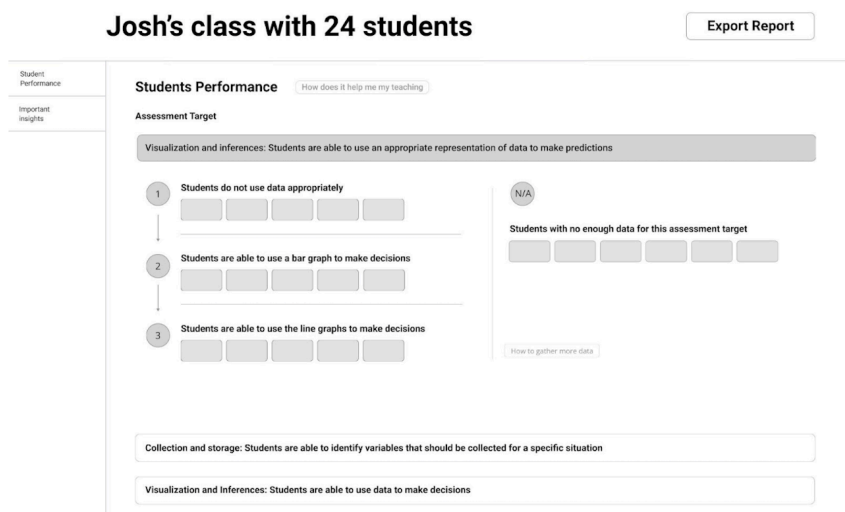
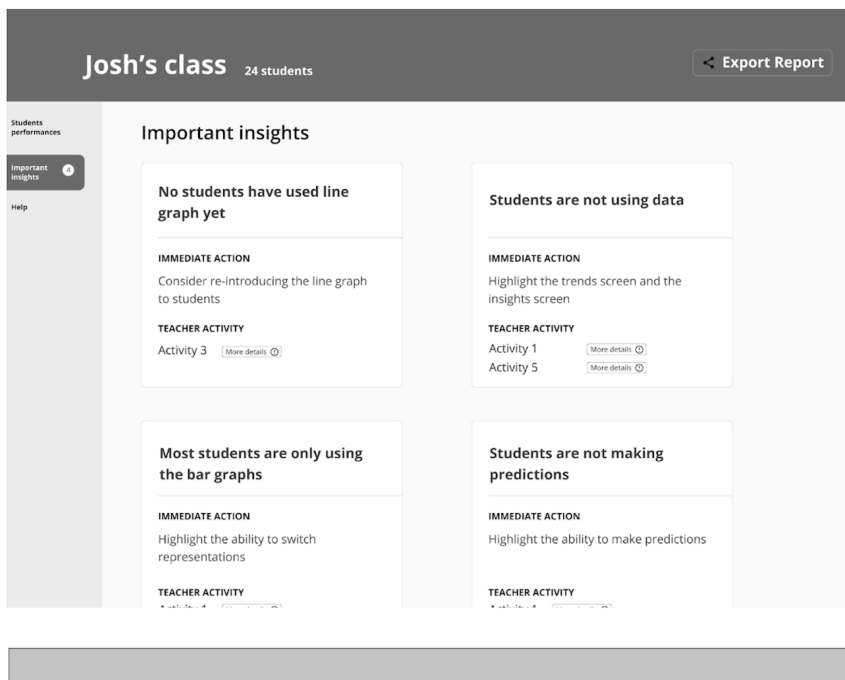


Figure 3. Prototype 2 The Important Insights page has three formative assessment outcomes and suggestions for next actions based on these formative assessments.

On this prototype, teachers appreciated the information organization – in particular the access to actionable information on student learning – but still felt the information was

“scattered” (“without a cohesive message”), suggesting that it still lacked a connection to acknowledged educational goals and standards. These participating teachers also indicated they wanted to compare statistics between students for engagement and track individual students’ learning progress, tracking how the students use data to inform their choices. They wanted the ability to explore this data more deeply.

Prototype 3. In the design of Prototype 3, the *Important Insights* page separates the insight on the left with various actions on the right providing the flexibility to grow vertically to scale (Figure 4). We refined the language on the II page and included additional pages containing details such as a pop-out box that compared student progress: one that compares students with each other for some key statistics, and another that compares students’ own progress over time (Figure 5).

Important insights


Many students have not visited the collection and storage screen.

[See students](#)

IMMEDIATE ACTION

1. Consider highlighting the collections/storage screen and encouraging students to interact with this screen
2. Discuss the importance of collecting and storing data. Include discussions of how to decide between different data to collect and how often data should be collected.

TEACHER ACTIVITY

[Activity 3](#) 



Some students are not using data to make decisions

[See students](#)

IMMEDIATE ACTION

1. Consider highlighting the value of using data to make decisions and demonstrating the insights screen (click 'Find trends') to encourage students to look at the data
2. Discuss how data can be used to decide a song trait.

TEACHER ACTIVITY

[Activity 1](#)  [Activity 5](#) 


Some students are only using the bar graphs

[See students](#)

IMMEDIATE ACTION

1. Demonstrate to students how to switch between graphs types, in particular, highlighting the line graph
2. Discuss the advantages of a line graph and the advantages of a bar graph.

TEACHER ACTIVITY

[Activity 5](#) 

Some students are not making predictions

[See students](#)

IMMEDIATE ACTION

1. Demonstrate the prediction function from the insights screen
2. Discuss how the data in the game can be used to make predictions

Students performance

How can it help my teaching?

Collection and storage: Students are able to identify variables that should be collected for a specific situation

level	Aware of collecting data	Students with no enough data for this assessment target
1	<div style="display: flex; justify-content: space-between; padding: 5px;"> Charlie Bob Ted Gary Molly Rachel </div>	<div style="display: flex; justify-content: space-between; padding: 5px;"> Ashe Jack Zoe Douglas Ela Ian </div> <p style="color: red; font-size: small;">How to get more information about all your students</p>
2	<div style="display: flex; justify-content: space-between; padding: 5px;"> Peter Xavier Yen Queen Olivia Vanassa </div>	
3	<div style="display: flex; justify-content: space-between; padding: 5px;"> Nick Wendy Sally Laura Felicity Happy </div>	

Visualization and inferences: Students are able to use an appropriate representation of data to make predictions

Visualization and Inferences: Students are able to use data to make decisions

Figure 4. The Important Insights & Student Performance Pages. The layout allows for the content to expand vertically, rather than horizontally, to accommodate class size.

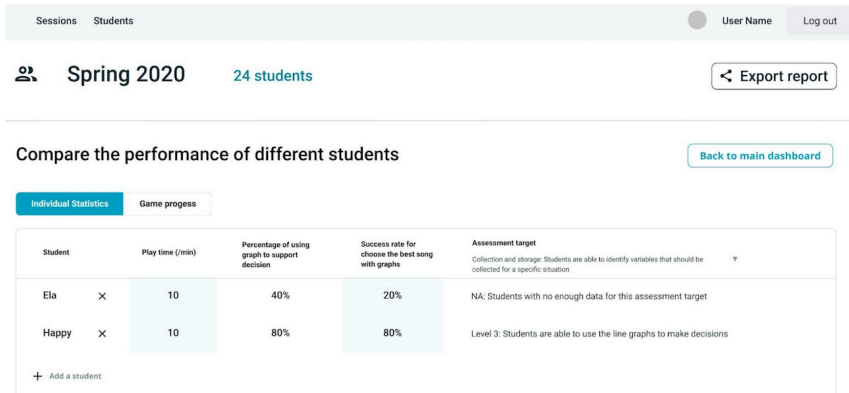


Figure 5. In the Student Performance Comparison page details are also visible that show comparative information on student activity as requested by interviewed teachers.

To evaluate these changes, participants were given a link to the prototype and shared their screens when exploring the dashboard, enabling us to evaluate the usability of the dashboard and its core features. This evaluation demonstrated actionable information and was deemed user friendly with a high System Usability Score (SUS – Bangor et al., 2008). The two shortcomings appeared to be: 1) the “view students” tab in the *Important Insights* screens lacks sufficient information to allow teachers to locate the problems of a specific student; and 2) some numeric statistics were not useful as they were unexplained and not tied to any action. In general, teachers used the dashboard actively, jumping back and forth between screens and carefully reading the texts and graphs on each.

When determining which page, *Important Insights* or *Student Performance*, would be the landing page, the teacher-participants held conflicting opinions, as did the research team. This lack of consensus suggested that navigating easily between these pages (which prototype 3 supported) was more important than choosing the “right” landing page. Additionally, teachers’ opinions on the landing page and the usefulness of specific features vary based on their experiences, current teaching practices, and skill sets, highlighting the importance of teacher

profiles that could shape adaptable content in designing this kind of dashboard.

Final design. In the final version of the *Important Insights* page (Figure 6), the teachers can find each insight numbered, with a message (left) and actions (right). The insight messages were defined by assessment experts and gave teachers indications of what learning may or may not be occurring. The right side has “Immediate Actions” which give teachers action tips that they can use to address those insights. The teaching activities refer to a set of systematic procedures devised to improve students’ graph reading ability with the game. An additional button (“See Students”) allows teachers to dive into detail, visualizing which students are struggling with which concepts (Figure 6). The future of this page design requires improving these texts so that teachers can more easily understand and interpret into actions.

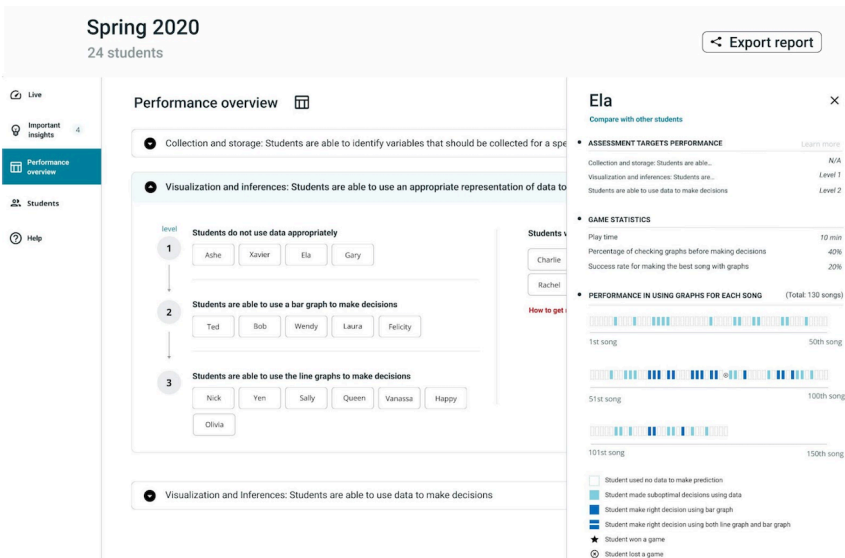
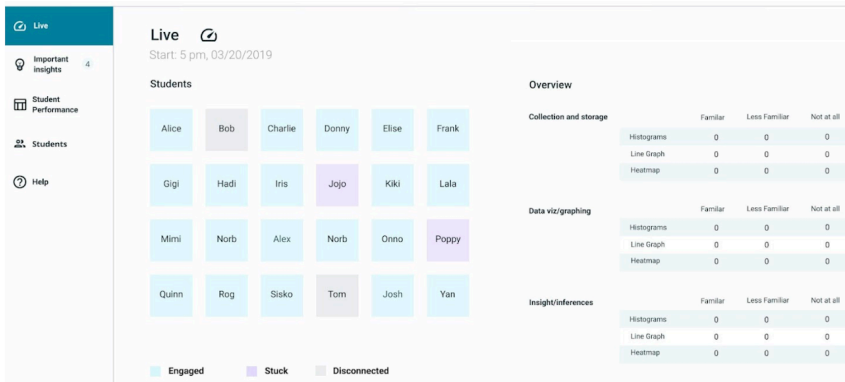


Figure 6. High fidelity mockups of the final design of the dashboard.

FINAL DESIGN

We summarized our final design with a list of generalizable design heuristics we call the Disciplinary Agnostic Dashboard Design (DADD) framework: **(01) Center teacher expertise; (02) Leverage Pedagogical Content Knowledge; (03) Use familiar and minimal UI patterns; (04) Connect behavior to actionable knowledge; and (05) Provide options to explore data.**

Early in the process we identified two needs that teachers have

for the dashboards: classroom management and lesson planning. These were then reflected in the two levels of data provided – individual student data and overall class performance. Our design process also corroborated many known findings across the field of developing teacher dashboards from different data sources – in this case providing a uniquely compiled list specifically for game-like formative assessment tools in classrooms. These decisions built on existing teacher knowledge about their classrooms and needs (DADD 01) and informed the basis for what kinds of feedback to design for (DADD 04).

Teachers' central role in productive implementation of games in classrooms is often hindered by a lack of tools to scaffold the games (Shah, 2019). Teachers were **receptive** to game based assessment tools, but needed **clear guides** connecting game content to standards (supporting DADD heuristics 02 and 04). At the same time, they were also clear that games looking substantially different from traditional assessment tools would not be considered suitable to replace scored tests. This supported our project's broader design goals, in choosing games specifically as *formative* assessment data sources and tools.

Teachers recognized **live data sources** and related dashboards as useful sources of information about students' behavior and engagement. For in-classroom use, they expect the utility and meaning of the data presented to them to be as immediately **actionable and usable** as possible. This matches a common perspective on the central use of dashboards (Brouns et al., 2015) – to provide data that can be consumed and used at a glance. Teachers' willingness to use complex visualizations is often overlooked, leading to the overemphasis of work on refining usability of dashboards instead of actual classroom utility (Ahn et al., 2019). Providing more "raw data" also leaves space for teachers to leverage their knowledge of students and their pedagogical expertise, creating their own nuanced interpretations, and acting appropriately. This example includes

explaining data as actionable feedback (DADD 04), while leaving space for minimally explained and openly interpretable data (DADD 01 and 05).

For the second use of these dashboards – as sources of information for lesson planning – teachers preferred general insights about learning outcomes and actionable items based on those insights. This format, intended as a reflective tool, provided explicit connections with standards, and situated learner activity around common mistakes, misconceptions, and points where students’ were stuck. These explicit connections exemplify heuristics 1 and 2 – data should be presented such that teachers can center their pedagogical content knowledge and other classroom expertise.

In this reflective format, teachers wanted both explanatory insights, as well as exploratory presentations that allowed them to explore the data with more depth. This enabled them to determine whether certain phenomena were specific challenges with **individual students** or **classwide** issues. Exploring the data might give them insights to see the problem in a new way or find insights to address the issue. **Exploring the data** also **builds trust** in the insights and suggestions of the dashboard. While teachers might use these deeper explorations infrequently, it gives the option to test or check generalized information.

A key takeaway is that dashboards are likely to be used by different teachers in different ways – or even in changing ways by the same teachers at different times. Teachers’ experience in the classroom, expertise in the subject area, confidence in technology all shaped the way they receive dashboards and concretized the value of providing both student performance data and learning insights data. Due to these variances, future work would explore the customization of dashboards based upon teacher profiles.

CONCLUSION

While work on educational dashboard design and research is productively pushing for richer work that centers teachers' needs as well as expertises (Ahn, et al., 2019), dashboards carry the potential to push classrooms to facilitate richer, multidisciplinary experiences for learners. These dashboards need to be parts of broader formative assessment systems and processes which present activities like *Beats Empire* to learners to elicit complex engagement and performances from learners which can map to multiple domains. Games are particularly productive for this approach given their broad potential to animate complex concepts and rhetorics (Bogost, 2008), and are also being increasingly accepted and implemented in classrooms.

In this work, we present a participatory design process where collaboration with teachers deeply informs the layout and content of our dashboard which can be used in classrooms in a spectrum of ways. An underexplored but critical aspect of educational dashboard research includes understanding the infrastructure needed in making it a sustained useful tool for teachers, including the initial training and an understanding of how it changes practice (Molenaar & Knoop-van Campen, 2018). While we believe that this design will be useful to teacher we recognize that more testing in active classrooms is needed to determine how well the dashboard supports productive use by teachers.

CHAPTER 8.

DESIGNING BRIDGING ACTIVITIES

SUMMARY

Teacher support is imperative when using Beats Empire effectively as part of formal classroom settings. Therefore, we present teacher supports designed to provide teachers with educative resources about data science concepts; practical information for classroom implementation; and discussion prompts and activity suggestions. We also describe examples of structured unplugged activities that bridge unstructured gameplay with the formal classroom setting, framing the game as an “object-to-think-with,” and grounding subsequent class discussion around similar data-based concepts.

INTRODUCTION

As we’ve shown in this book, educational games can act as useful formative assessment tools that help evaluate student understanding of target concepts in engaging and authentic contexts. However, teachers often need support with using games effectively as part of formal classroom settings, for

example, understanding how gameplay aligns with disciplinary content knowledge and practices, structuring classroom time around gameplay and other related activities, interpreting students' gameplay behavior, and leveraging and building upon students' implicit game-based behavior through related classroom activities and discussion prompts.

Games like *Beats Empire* can be very engaging for students while simultaneously revealing students' implicit understanding of the target concepts in an authentic context. We use the term *implicit understanding* to mean 'the ability to complete a task involving the concept without explaining it'. That said, literacy on specific concepts requires that students are also able to communicate those concepts to others and apply the ideas and practices on a variety of problems and in diverse contexts (Barchas-Lichtenstein et al., 2019). **Bridging activities** provide teachers with ways to connect game-based assessment with classroom content, thus translating students' implicit understanding into more explicit understanding. Studies by Asbell-Clark et al. (2020) show that when high school teachers bridged content in two physics games to classroom activities, students in bridging classes performed better on external post-tests, when accounting for pre-test scores, than in classes that only played the game or did not play the game at all.

Additionally, we found that gameplay in *Beats Empire* can be highly indicative of proficiency, however, lack of evidence of such actions does not necessarily indicate lack of data-related proficiency (Basu et al., 2020). Because students bring their own set of existing identities and literacies as data scientists, gamers, and music fans to the game, their choices in *Beats Empire* often reflect personal taste in music and general gaming strategies – as opposed to their abilities related to the educational content. As explained in other chapters in this book, we see that as a strength. However, when more explicit demonstrations of proficiency are needed, bridging activities can create opportunities for students

to articulate their understanding of data science concepts in context beyond the game.

In this chapter, we discuss teacher resources including bridging activities that we developed to assist middle school teachers in effectively using Beats Empire in formal classroom settings. In particular, we focus on how we might support teachers in leveraging and building on students' implicit game-based understandings. We present various examples of structured bridging activities that can be used after gameplay or between gameplay episodes to probe students' understanding of data science concepts. Finally, we show examples of teacher supports for facilitating reflection and classroom discussion around the game and bridging activities.

TEACHER SUPPORTS FOR USING BEATS EMPIRE IN CLASSROOM SETTINGS

We developed teacher resources in the form of a guide to support middle school teachers in their use of Beats Empire. The complete teacher guide includes all of the following information and resources.

1. An overview of the Beats Empire game
2. An overview of the middle school CS “Data and Analysis” concepts such as data collection, storage, visualization and transformation, and inference and models and alignment of the game with these concepts
3. Alignment between Beats Empire and target data skills for middle school students
4. Alignment of the game with other related standards in science (NGSS), math (CCSS), and social studies
5. Recommendations on how to introduce the game to students and suggestions on how to integrate the game into classroom activities depending on time available

6. Important definitions and vocabulary terms
7. Suggested post-gameplay classroom discussion prompts
8. A video showcasing how people in the music industry use data in their everyday work lives
9. Suggested unplugged bridging activities that probe deeper into concepts targeted in the game along with corresponding classroom discussion prompts

The goal of the teacher guide was to provide: a) educative resources on data science concepts and connections to the middle school computer science standards on “data and analysis”; b) practical information for introducing and using Beats Empire in the classroom and interpreting formative assessment information produced by Beats; c) prompts and activity suggestions to engage students in discussions around use of data concepts in non-game contexts; and d) resources to increase students’ awareness of CS and data science workforce opportunities.

Resources on Middle School “Data and Analysis” Concepts and Standards

While teachers implementing Beats Empire no doubt are knowledgeable about key data and data analysis concepts and practices, we found it useful to help make the relationship between the design of Beats Empire and domain standards explicit. The Data and Analysis strand of the K-12 CS Framework (K-12 Computer Science Framework, 2016) is broken up into four components: data collection, data storage, data transformation and visualization, and data inference and models. In the teacher guide, we provide teachers with an overview of the middle school “Data and Analysis” strand and show how Beats Empire aligns with the concepts and skills of this strand. For example, the teacher guide describes how the game emphasizes data collection concepts by allowing students

to decide what kinds of data to collect and how often to collect it. Also, what data is collected has implications for the amount and cost of storage required and what questions can be answered using the data. The teacher guide also describes the different data visualizations offered by Beats Empire (line graphs, bar graphs, and heat maps) depicting popularity of different types of songs in different locations; these visualizations inform students' decisions regarding types of artists to hire and songs to record. The teacher guide also introduces teachers to the assessment targets outlined in Chapter 5 and describes how various game features connect to those targets.

In addition to the CS Framework, the teacher guide highlights connections between Beats Empire and the current data related standards in other disciplines such as science, math, and social sciences. For example, the Next Generation Science Standards include practices such as *“Analyzing and interpreting data,”* *“Using Mathematics and computational thinking,”* *“Developing and using models,”* and *“Planning and carrying out investigations”* that Beats Empire designers intentionally aligned with gameplay activities. The teacher guide also elucidates the connections between Beats Empire and middle-school Common Core Math standards such as *“Represent and interpret data,”* *“Model with mathematics,”* and *“Construct viable arguments and critique the reasoning of others.”* These standards emphasize reasoning inductively about data and analyzing relations among quantities using tools such as tables, graphs, flowcharts, and formulae. Finally, the data-based decision-making process defining gameplay in Beats Empire also aligns well with the National Curriculum Standards for Social Science that include gathering and analyzing data to understand human behavior in relation to physical and cultural environment.

By acknowledging the range of disciplinary standards aligned with Beats Empire, and showing teachers how Beats Empire gameplay is related to these standards, the teacher guide makes

a strong argument for the value of this formative assessment game in multiple middle school classrooms. Likewise, we believe that making these connections apparent to teachers will prompt them to leverage their existing pedagogical knowledge to bridge gameplay to other classroom discussions and activities.

Recommendations for Introducing and Using Beats Empire in the Classroom

The teacher guide provides practical information for using Beats Empire as part of classroom instruction – how to introduce the game, how much class time to spend on the game, and how to integrate gameplay with other classroom activities. Teachers may choose to provide a brief introduction of the studio management context of the game but are encouraged to not explicitly discuss use of data during gameplay so that the game provides a fair assessment of students’ ability to notice and use available data for decision making.

As class time models can vary widely, we intentionally designed Beats Empire to be playable in both longer or short gameplay sessions. If teachers choose to use bridging activities as part of Beats Empire use in the classroom, we recommend using the game in four to six episodes with each gameplay episode lasting 15-20 minutes, and using bridging activities between gameplay sessions. These short gameplay sessions provide enough time for students’ actions in the game to produce meaningful data in the game dashboard (See Chapter 7) and to ensure students have a shared experience with data practices in the game to support interesting discussion of data-related concepts and real-world scenarios between gameplay sessions. If teachers have limited class time, we recommend the game be played just once for one class period (30-45 minutes), followed by one round of discussion.

Post-gameplay Classroom Discussion Prompts and Bridging Activities

A primary goal of the teacher guide is to provide teachers with discussion prompts and activities to support students as they make connections between experiences in Beats Empire and relevant data and analysis concepts in other classroom topics and real-world experiences.

Supporting discussions around vocabulary and gameplay. In addition to defining a few terms relevant to the game context (e.g. trends, borough, artists, etc), the teacher guide includes a list of important vocabulary terms and associated prompts that teachers can use for discussions with their students, preferably after students have played at least one round of the game. Examples include:

1. **Data** – Information collected for analysis or reference. In the game, the songs represent the data that is collected and analyzed based on their features. Prompts could include: (a) What do you think data is?, (b) Did you think of data while playing the music game?, (c) Can you think of an example of data from the music game?
2. **Metadata** – Metadata is data that provides information about other data. For example, the mood of a song, the length of a song, or the language of a song are all metadata about the song. Prompts could include: (a) Do you know what metadata is?, (b) Can you think of some examples of metadata in the music game?
3. **Variables** – Variables refer to factors or features that can vary or change. For example, when recording a song, things that can change about the song include the genre, the mood, the instrument, etc. Prompts could include: (a) What do you think a variable is?, (b) What are some variables in the context of purchasing market data?
4. **Query** – In the context of this game, a query represents

an inquiry into a data store or database to extract specific data entries that meet specified criteria. Prompts could include: (a) What are some queries you used?, (b) What are some of the criteria you specified in the queries?

5. Chart types and trends – A pattern of change in a process or a tendency of data points to move in a certain direction over time, generally represented by a graph. Prompts could include: (a) How many of you used a line graph in the game?, (b) How many used a heat map?, (c) How many used bar charts?, (d) Which graph type helped you notice a trend?, (e) Which graph did you find most useful during gameplay and why?, (f) Can you think of a scenario when you would want to use a heat map rather than a line graph?

The teacher guide also provides examples of post-gameplay classroom discussion prompts and suggested unplugged bridging activities that probe deeper into individual students' understanding of data related concepts targeted in the game. Some example prompts provided for 5-10 minute post-game discussions include:

1. The Brooklyn council is organizing a music festival and wants to invite the 3 most popular artists in Brooklyn in the last 6 months. How can they figure out who these 3 artists are?
2. The music studio wants to sign a famous new artist in one of the genres that has had the least number of followers in the last quarter. How can you help the studio manager figure out which genre that is?
3. Make a hint sheet for the next group of students who will play this game. What strategies will you recommend to succeed in the game?

Unplugged Activities that Bridge Gameplay with Learning in

Classroom Settings. The teacher guide includes a set of unplugged bridging activities aimed to allow teachers to probe deeper into the middle school data-related concepts targeted by Beats Empire. For each bridging activity, we designed a sample exemplar response and created corresponding classroom discussion prompts to support teachers in facilitating discussion around the activity.

Bridging activities can take many forms. For Beats Empire, we developed a set of eight unplugged bridging activities that each involve a scenario and a series of questions asking students to reason about data collection, processing, or inferences in the context of the scenario (see Table 1). Students can complete the activities either on their computer or on paper, as their completion does not require any digital tools. Questions do not always have a single correct answer, instead they are designed to elicit student reasoning using data. We designed the activities to be aligned with the same focal knowledge, skills, and abilities with which Beats Empire is aligned (Chapter 6). For each activity, we developed a student-facing and a teacher-facing version. The activities are designed for students to work on individually or in small groups before the teacher engages students in a whole class discussion based on their responses. Each activity is designed to take about 10-20 minutes of class time with students responding to questions individually or in small groups for 5-10 minutes and a whole class discussion continuing for around 5-10 minutes. The idea is that the game acts as a common object-to-think-with and helps ground subsequent class discussions around the activities involving similar data-based concepts (Holbert & Wilensky, 2019). Time permitting, teachers are recommended to interleave gameplay episodes with activities during class time. The teacher-facing version of the activities includes examples of desired student responses and sample discussion prompts to help the teacher facilitate discussions around the activity.

Some activities are directly tied to the game context while others

are based on different real-world contexts. Activities encourage students to think about data collection mechanisms to generate given data plots, automation of data collection for programming a music app, privacy concerns of data collection in a music app, the relative pros and cons of collecting data at different frequencies, resolving data stored in different formats, the amount of data needed to draw meaningful inferences, and how to draw inferences by combining information from multiple data representations.

Activity Name	Associated Assessment target	Approx. time needed
Identifying & Using Beats Empire Data	Knowledge of what data is and ability to identify examples of different types of data. Ability to interpret data visualizations, including a recognition of the different information provided by different data visualizations.	15 min
Human vs. Computer Activity	Ability to decide when a computer/computational power is necessary to process or understand a given dataset in a given context.	10 min
Data Collection and Storage Activity	Ability to identify variables that should be collected (and/or used) for a specific situation. Ability to develop a plan for collecting data including identifying what data should be collected and how often in order to generate a visualization and/or address a problem.	15 min
Data Collection Activity	Ability to identify variables that should be collected (and/or used) for a specific situation. Ability to use data to make decisions.	20 min
Graph Interpretation Activity	Ability to interpret data visualizations, including a recognition of the different information provided by different visualizations.	10 min
Data Sampling Activity	Ability to develop a plan for collecting data including identifying what data should be collected and how often in order to generate a visualization and/or address a problem. Knowledge of how much data is required to make decisions.	5 min
Data Formatting Activity	Ability to identify how data should be stored (e.g., format of the data).	15 min
Data Collection and Privacy Activity	Ability to identify variables that should be collected (and/or used) for a specific situation. Ability to describe privacy concerns related to the collection and storage of data.	10 min

Table 1. A set of post-gameplay activities designed to integrate gameplay with formal learning in classroom settings.

Figure 1 provides an example of the student-facing version of the “Data Collection” activity that measures students’ ability to identify variables that should be collected for a specific situation and ability to use the collected data to answer questions.

In Beats Empire, students can decide the frequency at which they want to collect different types of data such as popularity of songs by mood, topic, and genre. Collecting more data has storage implications, but besides that, students do not have to decide what types of data to collect or draw any explicit connections between the data they decide to collect and what they do with the collected data. The collected data is automatically presented in the form of data charts in the game. In the data collection activity (Figure 1), we allow students to think more explicitly about what data to collect and how to use that collected data to answer a given question.

Activity 4: Data Collection Activity

a. Desiree wants to create a NYCListens app to keep track of what songs people in New York City listen to and how many times they listen to each of these songs. The app will download data from a global music streaming service.

Below is a table with the possible types of data that Desiree can collect from the streaming service. Decide which types of data Desiree should use to determine the number of times songs were listened to for the entire month of April 2019 in New York City.

If Desiree should use the data:

- Select 'Use'
- Write the values of the data Desiree would use in the text box in the last column. If she should use all of the possible data then write "all".

If Desiree should not use the data

- Select Do not use
- Do not put anything in the text box in the last column

For example, for the data "The city in which the listener was located" -- select 'Use'. Under 'Specify which values of the data Desiree would use to answer her question', enter New York City.

	Data use		Specify which values of the data Desiree would use to answer her question
	Use	Do not use	
The city in which the listener was located			
Type of headphones the listener is using			
The day of the week of the listen (e.g., Friday)			
The date of the listen (e.g., January 5, 2019)			
The exact address of the listener			
The title and artist of the song a listener plays through the service			
The type of cell phone the listener owns			
A unique ID for each listener			
The name of the listener			

- b. Write instructions a computer can use to find the number of listeners in New York City for all songs on the streaming app for April 2020. You can write your instructions in a programming language, or you can write your instructions in regular English as a set of steps.

Figure 1. "Data Collection" Activity: Student-facing Version

Figure 2 (a and b) provide examples of resources provided to teachers to help facilitate classroom discussions around the data collection activity illustrated in Figure 1. For each activity, teachers are provided with an example of a desired response and some pointers for helping guide the classroom discussion. The

pointers are not meant to be prescriptive, but instead provide high level guidance on teacher-lead class discussions.

Desired student response:

- The city in which the listener was located – Use. Values: New York City
- Type of headphones the listener is using – Do not use
- The day of the week of the listen (e.g., Friday) – Do not use
- The date of the listen (e.g., January 5, 2019) – Use. Values: Month = April, Year = 2019, Date = anything
- The exact address of the listener – Do not use
- The title and artist of the song a listener plays through the service - Use. Values - ALL
- The type of cell phone the listener owns – Do not use
- A unique ID for each listener – Use. Values - ALL
- The name of the listener – Do not use

Discussion:

- i. **Selecting all data fields is not desirable since Desiree would need to pay for and store each data field she chooses to collect from the global music streaming app.**
- ii. **Listener id versus listener name** - Some students might want to choose the name of the listener. Impress upon students that different listeners might have the same name, hence having a unique listener ID will result in a more accurate count of number of listeners. Also, collecting listener names may result in privacy concerns.
- iii. **Listener city versus listener full address** – It is okay to choose either ‘listener city’ or ‘listener full address’ since the full address also contains the city. In order to decide whether the listener is from New York city, the only information required is the listener’s city. This warrants a discussion about how collecting the full address will require more storage and lead to privacy concerns, and hence collecting just the ‘listener city’ might be a better choice.
- iv. **Date of listen versus day of week of the listen** – To track listeners during April 2019, the date of listen will be useful to collect, but knowing which day of the week the song was listened to will not be helpful.
- v. **Title and artist of the song** – To track listeners for specific songs, it is important to collect data on the title and artist of the song in order to uniquely identify specific songs.
- vi. **Type of cell phone or headphone not relevant** - The type of cell phone or headphone used for listening will not provide any unique information about a listener or the song listened to. Hence, it makes little sense to collect such data.

Desired student response:

Desiree can write computer instructions to count number of listeners for each song she chooses as follows:

Repeat for all songs available in global steaming app:

- i. Choose song from global streaming app
- ii. IF ((song's date of listen contains 'April 2019') AND (song's listener city is 'New York City' or song's full address contains 'New York City') AND (song's title and artist matches that of the song Desiree chooses))
- iii. THEN, Increase count of listeners by 1

Discussion: Encourage students to describe a step-by-step set of instructions or algorithm for how the app must use the data it collects.

Students might describe an approach for counting listeners for a particular song and then filtering out listeners who are not from NYC or listens that do not happen in April 2019.

Highlight the importance of automating how the app uses the collected data, and how computers are powerful for searching for data and answering questions using data.

Figures 2. Classroom discussion facilitation prompts included in the teacher-facing version of the "Data Collection" unplugged activity for Parts A and B of the activity.

Increase Students' Awareness of CS and Data Science opportunities: A Video Showcasing Data Use in the Music Industry. Finally, the teacher guide also introduces resources that make real-world connections to the game and expose students to potential career options that integrate music, computation, and data analytics. When we first began exploring the music studio management game genre for the formative assessment game, we found many examples of companies supporting artists and labels to understand trends in the music industry. While students have many experiences with music and the music industry (see Chapter 2), we suspect they too may not be aware of how extensively data and data analysis skills are used throughout the industry. To fill this knowledge gap our project team recorded a video, "A Visit to Chartmetric," where leaders at a start-up explain why musical artists and their producers use data dashboards – such as, to plan a tour that promotes the artist in regions where they have followers. The gender and ethnically diverse Chartmetrics team explains what computational thinking with data looks like in the music industry, and why

computers are needed to collect, store, analyze, and make inferences from data. The data comes from streaming services, Wikipedia, social media, and other places where fans engage with artists. We encourage teachers to view the video with their students after a few gameplay episodes and facilitate class discussions around the video.

DISCUSSION

Designing bridging activities – activities that connect Beats Empire to more traditional forms of classroom assessment and instruction – involves making assumptions about what teachers might need and a careful analysis of what students might implicitly come to understand from the game. The activities emphasize surfacing those understandings. Even if students deeply understand the content in a game, it is the community – expert teachers, fellow learners, parents, friends – that enables them to connect it to their lives and to other academic subjects. In the next chapter, we show how a teacher used and personalized the bridging activities in her classroom.

However, what may be implicit in this chapter is exactly how we decided what game knowledge the activities would surface. What seems implicit but could be made explicit? Where do the “levers” come from? In our case, we started from the standards and found that they were often reasonably close to the gameplay. This was our intent, and it seemed successful enough to be a place to build. The game context around music management was evocative and extendable enough that we used it in many of our bridging activities, despite the explicit goal of connecting to content that is unrelated to the game. Having relatable contexts not tied to specific disciplinary content in science or social science help situate the bridging activities more broadly and make them accessible to a larger number of teachers and students.

CONCLUSION

This chapter describes how to support middle school teachers with using game-based assessments such as Beats Empire as part of their classroom instruction. Middle school teachers need support to develop and assess their students' data literacy skills in engaging and meaningful ways. To ensure teachers are able to leverage the full potential of the game-based assessment tool and dashboard, we believe it is useful to provide teachers with additional supports to organize instructional time around gameplay and associated activities, interpret and use information provided by the dashboard, and introduce students to real-life connections of the game.

Supplementing the use of Beats Empire with classroom activities and discussions can also help teachers to probe deeper into students' understanding of the data and analysis concepts of the game, and to support students as they make connections between these concepts and their experiences in other contexts outside the realm of the game. The game can act as a common object-to-think-with and help ground students' engagement with the activities and class discussions around shared experiences involving similar data-based concepts.

The teacher plays a critical role in leveraging the full potential of a game-based assessment tool by bridging the game with related classroom activities. In order to fully reap the benefits of Beats Empire and the associated classroom activities, it is critical to support teachers in making this connection. Helping teachers understand how the game and the associated activities relate to the same learning targets or FKSAs enables them to make stronger connections between the two and suggest bridging materials and activities on their own.

In the next chapter, we describe a small-scale study conducted in a middle school classroom in the Western United States with

one computer science teacher. We provide an account of how we prepared the teacher using a short professional development session where we walked the teacher through the teacher guide described in this chapter. We discuss what a classroom implementation of Beats Empire and related bridging activities might look like, how students engaged with the game and activities, how the teacher used the supports and resources we provided, and teacher insights and reflections on completion of the classroom implementation.

CHAPTER 9.

CLASSROOM IMPLEMENTATION

SUMMARY

In this chapter, we describe a week-long implementation of Beats Empire in a middle school CS classroom with one teacher and fifteen students. We provide details about how teacher's used bridging activities, and how classroom activities and students gameplay worked together to allow us to make inferences about students' evolving understanding of data.

INTRODUCTION

Beats Empire, and supporting classroom materials, were designed to serve as a tool for teachers to formatively assess what their students know and are able to do in regard to data and data analysis. Initial pilot studies focused solely on the game (e.g., game mechanics, data collection, usability, elicitation of the learning goals). From these studies we found that teachers found the game to be engaging and exciting for their students, but they were not always clear on how to best incorporate the game into their classroom instructions. This chapter focuses on a four-day

pilot study including both the game and the classroom activities that we designed to assist teachers with this integration. We first provide an overview of the pilot, and then describe what we learned about students using both student gameplay and their classroom activities. We end by discussing how these two different types of activities can support teachers in the classroom.

The pilot took place in a middle school in northern California. The pilot was done in a computer science classroom with fifteen seventh grade students (gender balanced) over the course of four days (about 45 minutes per day). The teacher who led the instruction – the students’ regular computer science teacher – has over ten years of teaching experience and previous experience as a computer scientist. They had already introduced students to the basics of data related concepts earlier in the school year.

Overview of the Pilot

Before the start of the study, the teacher participated in a two-hour professional development session. It included an overview of the teacher guide, a demonstration of the Beats Empire game introducing the key game mechanics and representations, and a walk through of the bridging classroom activities and their alignment to target data science concepts (see Chapter 8). Although the teacher guide provided recommendations on how to integrate the game and bridging activities into the classroom and we also provided guidance on these activities, the decision of which to implement and how to engage students with the game and activities was left up to the teacher.

During this study, we had two researchers observe the teacher and the classroom while the students played the game and engaged with the additional classroom materials. In addition to the observations, we collected students’ responses to the implemented bridging activities and a questionnaire created by

the teacher about the game – used as “exit tickets,” or a way to get student feedback about the day’s class before leaving the classroom. After the study, we conducted a teacher interview to receive teacher feedback on the game, the activities, and the other teacher supports we had designed. Lastly, we collected gameplay log data to examine gameplay behaviors.

Class Implementation

The general structure of each class period saw students spending about 15-20 minutes playing the game, 10 minutes working individually on paper/pencil activities related to their gameplay, and about 10-20 minutes engaging in teacher facilitated classroom discussions. This flow of activities matched our recommended flow. The teacher added exit tickets to obtain the students’ perspectives on their gameplay. Additionally, on day 3, the teacher modified the order, having students engage with the bridging activities first and then switch to gameplay. The teacher explained that she could use the gameplay as a “reward”, and it was easier to get kids to switch *to* playing the game than to switch *from* playing the game. Table 1 shows an overview of the four-day implementation.

Day	Class Activities
1	<ul style="list-style-type: none"> • Intro to game • Game play • Bridging Activity 5: Graph interpretation • Class discussion
2	<ul style="list-style-type: none"> • Review of previous day • Game play • Bridging Activity: Data Collection frequency comparison
3	<ul style="list-style-type: none"> • Bridging Activities 4 and 8: Data Collection and Data Collection and Privacy Activity • Gameplay • Exit ticket on gameplay features used, game performance, favorite aspects of the game
4	<ul style="list-style-type: none"> • Watch “Data analyst Music Industry” video • Class discussion • Exit ticket on gameplay strategies • Bridging Activity 7: Data Formatting

Table 1. Overview of the implementation of bridging activities. These class activities are similar to those presented in Table 1 of the Chapter 8.

Summary of Teacher Adaptations

While the teacher guide provided information on how to explain the game, it did not have guidance on how to frame the game related to the classroom activities (i.e., “why play this game”). For this purpose the teacher provided a presentation on “What do we do with data?” and told the students that for the activity they should “think like a data scientist.” During the gameplay sections, the teacher mainly let the kids play. The kids would often talk to each other, showing off the artist they hired, or checking on how high a song was on the chart. The teacher gave the students the activities as they were but did end up supporting some connections back to what they had done previously in the classroom. In particular, for one activity where students were asked to create an algorithm, the teacher elaborated this request by having them think about what the algorithm would look like in Python, as they had previously been learning Python programming.

The main modification that the teacher implemented was her use of her own exit tickets. The teacher wanted to capture students' thinking and the strategies they were using and so asked the students to talk about some of the aspects of their gameplay. From this the teacher could see who was using the predictions, who was thinking about the collection of data, and who was upgrading artists. Information was also captured about what students liked about the game – often focusing on making money and releasing good songs, and what they would want to add to the game – some suggestions were to make it more challenging, or the ability to keep going after you win.

TAKE-AWAYS FROM STUDENTS' GAMEPLAY AND CLASSROOM ACTIVITIES

While students showed understanding of how to use and interpret graphs in the activities, they differed in how they applied their understanding when playing the game. One of the main skills measured in Beats Empire is the ability to interpret the data visualizations such as bar and line graphs. These skills overlap with those measured by the 'Graph interpretation' activity. In review of student responses to the 'graph interpretation' activity, we found that almost all students were able to interpret and use both the bar and the line graphs (one student misinterpreted the line graph by reading it from right to left instead of left to right). When asked to explain their reasoning in making a decision using the line graph, many students provided an explanation that said the direction of trend line over time was increasing (e.g., "Also, it is going up the fastest in terms of how popular it is," "I would choose chill, because on the bar graph, chill songs are getting the most amount of listens, and on the line graph, chill listens are rising."). These responses indicated that students were able to interpret the line graphs correctly.

When it came to playing the game, analysis of students' gameplay data revealed that most students tended to stick to the bar graphs

if they used data at all when making decisions regarding which song to record. Figure 1 illustrates students' song-by-song data usage each day, and game wins and losses, and highlights how most students tended to stick to the bar graphs (as shown from the light green in Figure 1).

In Figure 1, a song was coded grey, and given a score of NA, if the player did not look at any data representation; blue, with a score of 0 if they looked at data but the data representation viewed was unable to give them information that led to the choices made when recording the song (i.e. they viewed a bar graph but predicted a song is trending up); light green, with a score of 1, if the student chose a characteristic that was most popular based on either the line or bar graph representation they viewed; and dark green, with a score of 1, if they picked a song that was trending up after looking at the line graph. The red represents when the student lost the game (ran out of money), the yellow when they won the game (at which point they started playing again) and the black represents when students stopped playing the game for at least two hours.

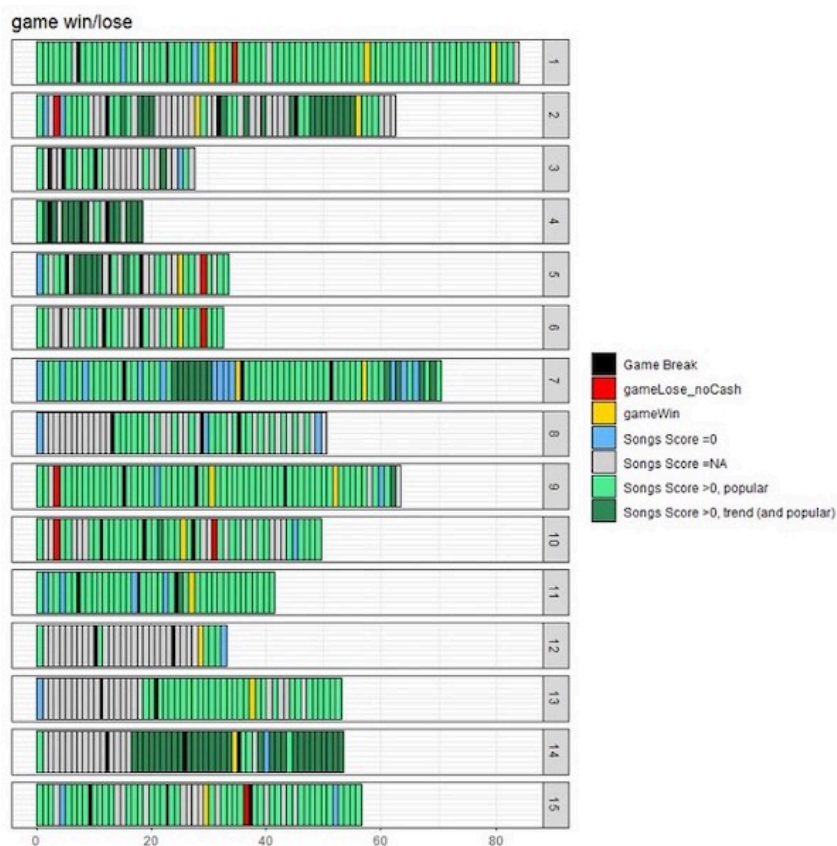


Figure 1. A representation indicating usage of data for each song they recorded. This figure shows that some students used data consistently (IDs 1, 7, 15), one student rarely used data (ID 12), and some students increased their use of data over time (IDs 8, 13 and 14). Some students also shifted in their use of data representations, for example ID 2 moved from mostly using bar graphs to mostly using line graphs for making data-based decisions. While students generally made decisions that matched the data they viewed (the green in the diagram) students did not always do this consistently (as shown by the blue boxes in the diagram).

From this graph, we can see that the conclusions drawn from the game were different from those indicated by the student responses from the activity alone, as from the game we see that students might not be comfortable with the line graphs, while the activities suggest that they are.

Both provide insights into students' use of data but it is only

when they are examined together that we get a more complete picture of whether students have acquired the desired knowledge and how they have chosen to use their knowledge in practice. One conclusion that we could draw is that while students conceptually knew how to interpret line graphs, in practice they tended to not use the line graphs and stuck with bar graphs. However, understanding when to use knowledge (i.e. when to use graphs) is an important part of a student's conceptual knowledge and as so should be something to include on assessments. Of course, there is still the possibility that for those students who didn't use the line graph, they may in fact, know how to use it but the game didn't provide enough motivation for them.

It is also worth noting that engaging with the activity did affect students' gameplay. We noticed that on the day after the students interacted with the Graph Interpretation activity (Day 2), seven students who had not previously accessed the line graph did so. Accessing line graphs is not equivalent to using data from line graphs to record songs, but it demonstrates how a classroom activity can help highlight and raise awareness about different game features and behaviors that students might not have noticed otherwise.

Understanding Tradeoffs of Data Collection Frequency

While students showed an understanding of the advantages and disadvantages of increasing the frequency at which data is collected, the game did not provide enough insight to determine if they could connect this understanding to real world data use. On Day 2 the classroom discussion centered around the 'Collection frequency comparison' activity, which compared the pros and cons of different data collection frequencies. Most students were able to reason that the more frequent the data collection, the more accurate the trends data would be, allowing for more accurate predictions. However, even after engaging with the activity,

students did not tend to make changes to the frequency at which data was collected in the game. This could be because students recognized that collecting data cost money (which we saw in the classroom activities) and they may not have seen an advantage to spending money to collect data. From the classroom conversations, as well as our previous think-alouds, students did not indicate that they needed more data to make decisions (see Chapter 6). In general students did not spend a lot of time in the collection and storage screen and very few students actually purchased more storage.

Determining What Data to Collect to Make a Data-Based Decision

While the game did not provide an opportunity for students to pick the data they were collecting, the activity was able to provide insights at how well students could identify what data needs to be collected to answer a given problem. Previous activities provided supplemental information about students abilities. Activities 2 and 3 measured assessment targets that were also measured in the game, and included images and questions from the game that allowed students to make direct connections to the game. The student responses, which provided reasoning behind their answers, afforded further insights on what students really understood about the assessment targets. Activity 4, on the other hand, provides information on an assessment target that was not assessed in the game.

The teacher set the context for this activity by announcing to the class, “Today, let’s think a little beyond [playing the game]. If you were designing a music app, what data would you collect?” While students were able to select appropriate data to use in their music app, they struggled with coming up with an algorithm to use the data to determine the number of listeners to a song in a particular month. The teacher supported students by linking the activity to what they had done in their class previously. Even with that support, some students still focused on the data they

needed to collect: “If the listener lives in NYC then it should collect the title and artist of the song. Also, Karen wants to know the data from the last month exactly.” Or they described the algorithm very generally: “She would use the unique ID for each listener to know how many listens the song got, and she would know the title of the song and the artist.” Only three students provided responses that were algorithmic and addressed the problem.

For example, one student who provided an algorithm included the following response: “1. Make sure the listener is in NYC 2. Make sure they listened in Apr 2019 3. Make sure the listener listened to the particular song 4. Make a virtual tally mark 5. Next listener id and repeat.”

From this activity, we learned that while students could identify data, they still needed support in figuring out how a computer could use this data to address a problem.

Students' View of Gameplay

Students viewed the game differently from regular classroom activities. During our class observations, we found that students enjoyed playing the game. We found from log data that students continued to play the game after they were out of school even when it was not asked of them, further indicating that they found the game to be a fun activity and not just something they had to engage with for school. When playing in school, however, they were highly engaged – often discussing the game and progress on the game with their classmates. For example, students would lean over to their neighbor and ask what ranking their song got, or they would announce if they won or lost the game. In contrast when students were working on the classroom activities they were quiet and did not interact with their classmates. Additionally, on the exit tickets, many students stated that their favorite part of the game was making money and recording songs, with one student stating that their favorite part was “the

ability to judge the trends of the music.” The teacher-led discussions of the classroom activities did engage students more than solo work, but the class was still subdued and did not provide a lot of unprompted discussion.

CONCLUSION

Overall, we found that both the game and the activities were useful tools in the classroom and can be used to obtain complementary information about students. While the classroom activities provide snapshots of the information that students know, the game-based activities can provide information on how students use this knowledge and how this use changes over time.

Apart from the feedback provided by the game, we also found that the teacher appreciated the music studio context of the game and the way it encouraged students to engage with the concepts of data and analysis. The game provided a unifying context around which the class could have discussions about the use of data in the real world. The activities were also useful in bringing in additional skills, for example having students engage in written communication and measuring students’ knowledge of data privacy.

One area in which further improvement could be made would be to add in additional links between the game and the classroom activities, such as allowing students to download data that showed information about each of the songs they recorded. From this data students could better analyze their own gameplay, and determine what aspects led to higher ranked songs.

Teachers can benefit from having a well-rounded picture of what their students know and are able to do. This can help guide instruction and highlight where students could use support as well as what they are doing well. Having multiple types of assessments can help provide a comprehensive view of students

if the different types of assessments are designed to measure different aspects of student learning. Using a game-based assessment along with additional classroom materials that are designed to complement the information provided through students' gameplay can help teachers obtain this picture of their students, and provide an interactive environment for students to actively engage with the concepts they are learning in the classroom.

PART III.

**PLAYLIST 3: BEATS EMPIRE
RETROSPECTIVE**

CHAPTER 10.

BEATS EMPIRE HITS AND MISSES

SUMMARY

The design intention of Beats Empire was to create an assessment of computational thinking for urban middle school students by designing a game with a music industry context. The team considered how the game design might be authentic, playful, and relevant to students in New York City with whom it would be tested. The three concepts of authentic, playful, and culturally relevant learning each have a history in the learning sciences literature. A reflective look at Beats Empire considers both hits and misses relative to these topics in published learning science theory. By looking at the hits, we can reflect on what worked in the design of Beats Empire and therefore, what might be carried forward or generalized to other playful games for computational thinking. In looking at the misses, we can consider further improvements to the genre of playful assessments of computational thinking. Based on this reflection, we suggest that future research with this game (or similar ones) could explore the synergies among authentic, playful, and culturally-relevant approaches more deeply. For example, by

going more deeply into authentic technological, social, and cultural dimensions of the music industry context. Designs could also double down on playfulness to expand from in-game activities to other classroom activities. Aspirational goals for a culturally responsive assessment—for example, measuring not only a student’s academic success but also growth of a positive identity—could be addressed also in further design-based research.

INTRODUCTION

As described throughout this volume, *Beats Empire* is a playful assessment of computational thinking. *Beats Empire* applies the management game genre in order to invite each student to play the role of a decision maker in the music industry. The student manages the songs recorded and released by their fictional music artists to increase the artists’ followers and their income.

On its surface, the conceit of the game connects with three key design concepts: authentic, playful, and culturally-relevant. *Beats Empire* may be considered “authentic” in the sense that today’s music artists, producers, and managers regularly analyze data to promote their music and artists. *Beats Empire* may be aligned with “playfulness” in that students can choose their own goals and explore freely. *Beats Empire* may be accepted as “culturally relevant” for middle school students in New York City because music is important to youth culture; students thus have new opportunities to show what they know and can do in relationship to the community assets related to music and the music industry.

Skepticism is warranted, however. In reference to learning

technology designs, the term “authentic” has been used in many ways. Additionally, “playful” and “assessment” are odd bedfellows. For example, because assessment often requires some degree of external control in order to achieve validity, this control can limit playful choice and exploration. Finally, claims of “cultural relevance” can be superficial. They may fall short of the drive to center learning or assessment in students’ cultural competence or can fall similarly short of examining multiple dimensions of success – which might include students’ development of academic knowledge, positive STEM identity, and abilities to identify and address inequity.

The previous chapters in this volume provided two kinds of insight: about how the game was designed and what we learned from research. For example, the third chapter shared the intention to design a game where the subject matter content and the gameplay were closely intertwined, where students who have considerable choice about how to play, and where careful attention to the game’s theme (the music industry) and students’ relationships to music might make the design culturally responsive. In this and other chapters, tensions, challenges, and uncertainties in the design process are addressed, illustrating an overall point that design of playful assessments is an area in which the field has a continuing need to explore alternatives and build more robust knowledge and recommendations. With regard to research, several types of research were conducted. Notably, the cognitive labs “revealed how each player incorporated their own experiences and background knowledge of music, and used them as assets to use the data and make decisions” (see Chapter 6). The intent of each research effort was formative; correspondingly, each research study revealed ways in which the current Beats Empire design was successful and ways in which improvements are needed.

Building on these design and research chapters, this chapter reflects more broadly on the design of the game and what was

learned in the context of these three relevant theoretical frameworks and asks “where are the hits or misses?”

In the music industry, an intended “hit” often appeared on the first side of a vinyl record, with another song on the back or “B-side.” Sometimes b-sides were bigger successes than the intended hits. In that spirit, the paper will discuss hits and b-sides. The intent is to understand more clearly from the “hits” what worked about the design. More specifically, to identify key design elements that could be further built upon or generalized as this or other teams move forward in building games for learning and assessing computational thinking. The intent is also to understand from the “b-sides” where future design work could dig in more deeply. The overall argument is that Beats Empire is a promising start towards integrating “authentic,” “playful assessment,” and “culturally responsive” in a coherent design; and yet, the implications may be even more powerful if future work could engage these three relevant frameworks more deeply—being more attentive to the full scope of each dimension and more committed to intentionally leveraging the intersections across the dimensions.

AUTHENTICITY

Recent definitions of authenticity include several components: they emphasize opportunities to do real world tasks with realistic resources or tools; they maintain that world tasks are not well-defined and require investigation over time; and they involve taking on a socially recognized role or roles (Herrington, Oliver & Reeves, 2003; Herrington, Parker, & Boase-Jelinek, 2014). Authentic tasks “engage learners as professionals” (Elliot, 2007, p. 34). Herrington, Reeves and Oliver (2010, p. 17) provide this comprehensive list of characteristics of authentic, technology-enhanced learning:

1. Provide authentic contexts that reflect the way the

knowledge will be used in real life

2. Provide authentic activities
3. Provide access to expert performances and the modeling of processes
4. Provide multiple roles and perspectives
5. Support collaborative construction of knowledge
6. Promote reflection to enable abstractions to be formed
7. Promote articulation to enable tacit knowledge to be made explicit
8. Provide coaching and scaffolding by the teacher at critical times
9. Provide for authentic assessment of learning within the tasks

As a design goal, authenticity has a long history in the learning sciences. For example, Means and Olson (1994) noted that technology was being used ineffectively for short exercises that supplement didactic instruction, but that in some more forward-looking applications “students are challenged with complex, authentic tasks, and reformers are pushing for lengthy multidisciplinary projects, cooperative learning groups, flexible scheduling, and authentic assessments.” (p. 16). Means and Olson saw three reasons why technology could support authenticity: complex assignments could become more feasible; advanced topics might now have earlier entry points; tools could scaffold what students can do.

One seminal project, the Fifth Dimension, used technology and games to create a unique after-school space for student learning. Relevant to authenticity, we note that in their analysis of the Fifth Dimension as a “Construction Zone,” Newman, Griffin and Cole (1989) conceptualized the role of technology as helping to create a Zone of Proximal Development (ZPD, discussed in Wertsch,

1985). Whereas placing young learners in the full complexity of a real world situation would be inappropriate, technology could be supportive of designing a “zone” where complex, open-ended assignments would be more feasible; advanced skills and knowledge might develop; and tools might scaffold what students can do, enabling them to engage in important social and cultural practice in ways that are authentic and yet manageable. That is, technology can enhance authenticity by making a later socio-cultural practice more accessible to students now, while preserving and enhancing the opportunity for the students to engage in meaningful learning. The ZPD-oriented perspective is somewhat more learner-centered—the ZPD-oriented view asks us to consider also whether the technology *enables a student to participate in meaningful socio-cultural practice that would otherwise be inaccessible*.

Authenticity: Hits

The design of Beats Empire was anchored in the K-12 Computer Science Framework (2016). The data strand of this framework has four components: collection; storage; visualization and transformation; and modeling and inference. These framework components offer a baseline for what computer scientists consider to be the signature authentic activities related to data and its analysis. As described throughout this volume, the team elaborated what these components mean for middle school students and then sought to embed them as activities in Beats Empire.

The design is best explained somewhat out-of-order. In order to promote an artist’s career in Beats Empire, the student in the role of manager uses data visualization and transformation to examine trends in the popularity of songs in a simulated city. In a nod to realism, different songs are popular in different regions of the city. The data visualizations allow looking at the data in different ways (regional heatmaps, comparisons of current

popularity, trends) and analyzing the attributes of songs that underlie the patterns (such as the “theme” or “mood” of the song). By investigating data, students can make inferences about what is driving popularity and form a mental model of the characteristics of songs their artist might release. Back in the recording studio, the student can adjust the title, mood, and theme of the songs to be recorded and decide in which neighborhoods they should release the songs. Further, as the game develops, they can decide what data they would like to acquire (“collect”) to enable making even better choices. Two components—visualization and inference are well-addressed in Beats Empire through a meaningful cycle of gameplay. Additionally, making choices about what data to collect can be important to a student’s game strategy (although the team acknowledges that opportunities to achieve the “collect” component of the framework could be elaborated in the game). One component of the framework, storage, proved even more difficult, as will be discussed in “Misses” below.

We also conducted a small set of industry interviews to examine the realism of Beats Empire relative to socio-cultural practices in the music industry (Roschelle, 2019). We found that the Beats Empire game was strikingly real in the eyes of the informants who looked at the game. According to our informants, today’s artists, producers, and managers do use data when producing and releasing new songs. In addition to these interviews, we analyzed documents on the web and found that big companies like Spotify and Pandora as well as smaller startups sell data analysis services to artistic teams for the purpose of guiding what they record next. As in the game, today’s music producers do consider the thematic content and mood of the song, as well as other features like the tempo, instrumentation, range of dynamics, etc., and then adjust which songs are recorded and released to capitalize on trends. They conduct related analyses when deciding where to tour or in what regions to promote

music. Further, as with the Beats Empire “collection” feature, teams make choices about which data sets to buy and integrate to support these decisions. We learned, for example, that page views of an artist’s Wikipedia page can be an early signal of recognition in a new region, and artists might decide to tour where there is an uptick in these views. We also learned that acquiring and merging data is expensive, necessitating tough choices when it comes to data collection and storage.

Further, experiences with Beats Empire in classrooms (reported in other chapters in this volume), suggest that a third very important stakeholder saw Beats Empire as authentic: the students. We readily observed that students engaged with Beats Empire as an authentic game. They understood the game as representing a familiar music industry context; for example, they recognized some of the artist names in the game as similar to real world artists with whom they were familiar. They also recognized the game’s major mechanic; using data to improve artists’ followers and sales is meaningful and realistic. Overall, a clear hit of Beats Empire is that it does let students engage in a social-cultural practice—using data to manage a musician’s growing success—that would otherwise be inaccessible to them. It reflects a real-world practice via authentic processes of making decisions about what music to record by using data.

Authenticity: B-sides

With regard to the CS K-12 framework, Beats Empire was not particularly developed with regard to exploring the concept of “data storage.” The team considered adding storylines that might enhance opportunities to assess what students know and can do with regard to data storage, but found these storylines could detract from the coherence of the gameplay experience. Given that the game would only last a short time, taking detours to address “storage” could weaken the continuity of the game experience and detract from the assessment goals.

For example, students might boost popularity for their artist by designing audio and video playlists that could be downloaded to a mobile device and viewed without using network bandwidth (no doubt, students have experience with the limits of their household data plans). Students could think about different storage formats and space tradeoffs that might affect their listener's ability to store music and video on their mobile devices. However, the core game loop did not have a playlist concept, nor did it have any sense of the data storage requirements of songs or videos or the capabilities of individual listener devices. Adding all this information might have been quite distracting from other assessment purposes, such as those related to analyzing data and making inferences.

Interestingly, the team later learned that the missing "playlist" concept was one area in which the game fell short on authenticity relative to industry norms. When we interviewed experts in the industry that provides data to artist management teams, they reported that getting a song into a popular playlist is an important way to boost an artist's career and a good amount of data analysis is aimed at finding the playlists that would both fit the artist's repertoire and grow their audience. In future game development, it seems possible that playlists would be something that middle school students would understand and that could bring more realism to the game—and as previously mentioned, it could help with the challenge of creating realistic activities for middle school students around data storage.

Additional b-sides relevant to authenticity can be observed by reviewing the nine characteristics of authenticity described above. We note especially that the game lacks a concept of mentors and of observing expert performances, and that the game could do more around supporting reflection and making implicit knowledge more explicit. One reflection is that there is a missed opportunity to build into the game a deeper sense of cognitive apprenticeship (Brown, Collins & Duguid, 1989).

Herrington, Reeves and Oliver (2010) discuss a foundational relationship between “authenticity” and the learning sciences concept of cognitive apprenticeship. Cognitive apprenticeship involves learning with mentoring and coaching from a mentor, coach, or teacher. For example, in the Fifth Dimension, Michael Cole and colleagues designed after school clubs in which undergraduates worked with younger students, mentoring them on important socio-cultural problem solving practices. Access to expert performances and to mentoring is limited in Beats Empire. Further, it presents to students more as an individual learning experience than as a social or collaborative learning opportunity. Beats Empire has one predominant role as the talent manager, whereas in real life, the industry divides the work around data, songs, and distribution of music among multiple roles. It would be plausible in an elaborated game, for example, to have roles around the same music that are more oriented to preparing and organizing data, to making decisions with the talent, and to distribution decisions (e.g. placement in playlists, where to tour, etc.).

When we made a site visit to watch experts who develop data analysis tools for the music industry, we actively saw cognitive apprenticeships underway among the employees on the team as they examined data displays and discussed what they might mean and how to better show information that supports actionable insights. Along these lines, we saw a missed opportunity for career education. The industry people we spoke with were diverse in terms of gender, ethnicity, and race. They described this industry as connected to their own passions with regard to music, and also relatively open to their employment and growth. They were able to describe how they became involved, their own mentoring experiences, and how being in this industry supported their own positive sense of cultural identity. Put simply, they loved building tools that supported the kinds of music and the artists who they valued. We see the cluster of

concepts related to career education and growth of students' positive cultural identity as an area that would be well worth developing in the future.

Authenticity: Reflection

Overall, we see the music industry scenario at the heart of Beats Empire is rich, with many possibilities for elaboration and extended engagement in ways that would create more authentic learning experiences. The b-sides around the “playlist” concept and opportunities to incorporate a cognitive apprenticeship overlay into Beat Empire could make the game both more authentic and further reaching in its formative assessment capabilities. In the next two sections, we will further reflect on how exploring synergies among the b-sides of authenticity, playfulness, and cultural relevance will be important going forward.

PLAYFUL ASSESSMENT

As was the case in the preceding discussion of authenticity, the concept of playful learning has a long track record in education. Piaget (quoted in Elkind, 2008, p. 3) wrote “Play is the answer to the question ‘How does anything new come about?’” In his theory of development, Piaget held that children learn by exploring and experimenting, a process which can lead to cognitive dissonance between their intuitions and how the world responds, and to transformation of cognitive structures to achieve better alignment. Plass et al. (2015, p. 287) elaborated. “Playful learning can be defined as an activity by the learner, aimed at the construction of a mental model (a coherent representation of the information in memory), that is designed to include one or more elements of games for the purpose of enhancing the learning process.”

Play is not just cognitive, but also social—students often engage in playful learning with peers and with mentors. Returning to the

example of the Fifth Dimension, we see a playful construction zone that fits not only Piaget's enthusiasm for play, but also the Vygotsky notion of a zone of proximal development cited earlier. When students played in the Fifth Dimension, they displayed aspects of their knowledge and skill to peers and mentors, and these could become zones for assessment and learning where the game mediated a reflection about more general knowledge or skill. Further, the Fifth Dimension had a concept of a wizard who directed students to new challenges, resolved disputes, and provided strategic guidance (Cole & Distributed Literacy Consortium, 2006). The social dimension also resonates with the research of Gee (2007) who found that learning and assessment occurred not only in a game, but also in the discussions around it. Finally, we observe that play is also affective. The suspension of reality in game-like environments can create safety in taking risks, learning from mistakes, and reflecting on one's own learning (Prensky, 2006).

The Beats Empire project emphasized playful *assessment*, a somewhat newer design genre. Seminal work in this genre was led by Val Shute (2001), who coined the term "stealth assessment" for assessments that were embedded in typical learning activities rather than occurring as clearly demarcated resources and occasions for measuring skill and knowledge, i.e. tests. The vehicle for developing the concept of stealth assessment has been designing learning games, and hence stealth assessment and playful or game-based assessment are closely interrelated. Shute (2011) describes "a quiet, yet powerful process by which learner performance data are continuously gathered during the course of playing/learning and inferences are made about the level of relevant competencies. Inferences on competency states are stored in a dynamic model of the learner. Stealth assessment is intended to support learning and maintain flow..." (p. 504).

Shute's work on Newton's Playground (Kim et al., 2016) sets a high bar for what is possible in playful assessment. In this game,

students seek to get a ball to roll, bounce, fly, and otherwise move from a start point to a target. To do so, they sketch physical structures like ramps, floors, and walls that will interact with the ball (through Newton's Laws) to change its trajectory. As a basis for assessment, Kim et al. (2017) adopts Evidence-Centered Design as a means to elaborate the connections about target competencies (to be assessed), situations in which those competencies could be elicited and observed (a game), and rules of evidence by which inferences could be made about individuals. Shute et al. (2020) describe the many intensive cycles of iterative design research that were necessary to achieve a game that could yield valuable and valid assessment information. A key takeaway message from her team's work is how much disciplined and iterative work is required to create a playful assessment. Research confirmed that when students engage in gameplay and assessment, their overall learning can be improved (Shute et al., 2020).

Playful: Hits

The A-side or "hit" of Beats Empire was clearly that it succeeds as an enjoyable, serious game for students who play it. We observed students spontaneously competing to see who could acquire the most money and pairs of students sharing headphones to appreciate one another's recently released hit.

In considering our team's experience with Beats Empire in classrooms, one can immediately see aspects relevant to Piaget come to the fore. Students readily explored the game without much guidance; they were experimental and learned from their experience; discussion with students suggests they were building relevant mental models (see Chapter 6). There was sometimes dissonance between students' intuitions or impulses about managing an artist's recordings and what could be understood by analyzing data. Students experienced challenges in relating

the less familiar forms of plots to inferences they could make to advance their gameplay.

We also see a fit to Shute’s seminal work in several ways. As with Newton’s Playground, Evidence-Centered Design was used to define the competencies that Beats Empire would seek to assess, as well as how a students’ knowledge could be observed and evaluated through gameplay (see Chapters 2, 4, & 5) We also observed that Beats Empire was playful as experienced by students, (See Chapters 2 & 6) and that a useful data dashboard could result for teachers (see Chapter 7). Like Newton’s Playground, Beats Empire appeared to function as a stealth assessment, where students generated assessment information without feeling they were “taking a test.”

Yet, Beats Empire was more limited than the games that Shute described: gameplay with Beats Empire was typically short— just 1-2 sessions of 20-40 minutes each. Based on feedback from teachers regarding the duration of classroom time they could reasonably allocate to a playful assessment, the design team intentionally designed Beats Empire for about one hour of total play time. In this available time, the number of competencies that could be assessed was small and the degree of precision of each assessment did not reach the team’s aspirations. In contrast, Shute’s team engaged students for longer time periods and could explore concepts more deeply and with greater precision. Shute also described more extended series of design iterations than was possible within the scope of the Beats Empire project; more iterations could have been useful to refine the Beats Empire playful assessment design.

Playful: B-sides

The most important “b-side” appears to be a design strategy that emerged in the product. In the course of reflecting on the tensions between playfulness and assessment early in the project, the team recognized and decided that their overall goal was a

formative assessment *system* that would include two complementary parts (see Grover, 2021 for more on the wisdom of focusing on assessment *systems* for computational thinking). The team decided that the game itself would be more student-centered, emphasizing elements that motivated gameplay while also collecting assessment data for a dashboard. This would be complemented by teacher-led discussions (leveraging the dashboard) which could further probe what students know and can do on the four concepts of collection, storage, analysis, and inference. Both the game itself and the teacher discussion guides referenced the context, storyline, and game mechanic of Beats Empire, but each had different affordances for assessment. (See Chapter 7 for more discussion of the dashboard, and Chapter 8 regarding the teacher discussion guides).

One nascent hit is that the teacher-led discussions were able to be playful as well, by—for example—inviting students to think about “what if” questions that went beyond their experience in the game. Hence, we realized that a *playful assessment system* can include both game-like and discussion-like aspects, each of which references a common context that engages students’ imaginations, allows them to construct alternatives, and invites a broader sense of evaluation beyond simply being “right” or “wrong.” In this way, play need not be restricted to in-game play, but transcends the game.

On a pragmatic level, a focus on the system allowed us to continue iterating on the overall quality of Beats Empire as a playful assessment without the cost of updating the game itself—iteration could continue in teacher discussion guides, and these iterations were relatively inexpensive and quick to design and test. Also pragmatically speaking, we see the use of Beats Empire as fitting a long tradition of play and assessment in education. Teachers have long given students items to play with—physical manipulatives for example. Teachers may observe how students engage with these manipulatives to get a sense

of what they know and can do, and also, at some point, may pose a particular task or question to better probe students' understanding. This mix between exploratory playful engagement and probing for understanding is recognizable and familiar to most teachers.

We also see potential b-sides, relative to Fifth Dimension, that resonate with our earlier discussion in the authenticity section. The two designed modes of Beats Empire were individual student gameplay and teacher-led discussion. There was less explicit learning or assessment design for either small group discussion groups or for teachers to interact with students as mentors during gameplay (although, with gaming and forthcoming AI technologies it would also be possible to design in-game mentor characters). Yet the playfulness of games has a natural synergy with playful student discussion and with mentoring elements. The concept of a playful classroom experience could, for example, expand into role-play exercises in which students articulate different possible strategies for using data to grow their artists' reputation—and decide to work as a team to see which strategies are effective in the game. Again, we see this “b-side” as related to the limited time available to design and test Beats Empire—there was enough work just to define the game, its dashboard, and relevant teacher-led discussion activities.

Playful: Reflection

It is an important success that Beats Empire was accepted as a game in the eyes of its audience; students are tough critics of games that do not meet their standards. The finding that a combination of the management game genre with a music industry context is believable and enjoyable can be built upon as the game is refined.

Yet, we believe the tensions between assessment and game design persist, and thus we suspect that the right focus of design going

forward is on a *playful assessment system*, and not just on the game. In this systems approach, the game can be playful in the obvious ways that games are intrinsically playful. There is much good in this—students enjoyed playing the game and the opportunities therein to take risks and explore freely. There can also be cognitive and social good. When students are in the high state of flow (i.e. Csikszentmihalyi, 1990) as Shute describes—an experience marked by feelings of immersion, being challenging, the ability to take action, and a struggle leading to good outcomes—they can fully engage their cognitive powers in ways that more mundane school tasks do not elicit. Further, since our assessment goal was to inform teacher instructional decision making, and not to give each student an individual score, it is possible for the game to be social. This can be favorable not only for student learning, but also for the teacher who can observe what students know and can do through talk and interaction among students.

As Beats Empire demonstrates, however, prioritizing play does not mean giving up on assessment. As seen elsewhere in this book, dashboards can provide indicators to teachers about possible strengths and shortcomings of student knowledge. In addition, teachers may interact with students while they play Beats Empire or probe students' thinking in discussions afterwards. Thus the time spent in gameplay is not only useful for activating and eliciting each student's minds-on cognitive engagement, but also can produce a first stage of assessment insight. Building on this, the teacher-led portion can be more playful because teachers can contextualize their discussion in the game. A constant struggle in teacher-led discussions is to elicit more than short utterances from students. But after recently having played a game, we observed that students are better able and willing to share longer, more elaborate thoughts, such as possible rationales, strategies, concepts, and approaches. Thus, the "playfulness" embodied in the game can carry forward into

a more animated teacher-led assessment discussion. Finally, formative assessment implies an involved teacher who is not only obtaining assessment information, but is also using it for instructional decision making. Getting the teacher more active in the overall playful assessment cycle—for example, taking the role of mentoring the students or playing the “wizard” (as in Fifth Dimension) is likely to yield better overall formative assessment. As we move into the next section, we’ll seek to develop how an even stronger focus on authenticity and play might result in a game with enhanced cultural relevance.

CULTURALLY-RELEVANT

Ladson-Billings (1995) introduced the term “culturally relevant pedagogy” a quarter-century ago. Underlying the framework are three central goals around culturally relevant teaching: (1) prioritize students’ academic success, (2) help students develop positive cultural identities, and (3) support students’ ability to both identify and critique current social inequalities. Broadly speaking, culturally relevant pedagogies aim to dismantle a deficit-oriented view of Black and minoritized students’ skills and knowledge; instead, these pedagogies seek to identify and build on assets that exist in students’ experience, community and culture (Gay, 2010). Gay provides a prescription for teaching practice that would fully embody these ideas. Gay’s recommended teaching practices involves changes to many dimensions of teaching, such as how teachers care for students, communicate with students, choose or develop curriculum, and enact instructions. Going beyond the idea of culturally relevant, Paris and Alim (2017) describe “culturally sustaining” pedagogies; these pedagogies seek to strengthen both the students’ traditional cultures and their emerging, lived cultures; culturally sustaining pedagogies seek to develop an enriched and positive sense of identity for each student.

The majority of research around culturally-relevant and

culturally-sustaining concepts appears to be related to curriculum and pedagogy, with less attention paid to culturally relevant assessment. Nonetheless shortly after the 1995 publication of Ladson-Billing's work, there was a special issue of the *Journal of Negro Education* on "Assessment in the Context of Culture and Pedagogy." The intent of the special issue was "...to provide a critical discourse that informs discussions and explores strategies relative to the potential of culturally responsive, performance-based assessments to assess students and teachers of color" (Stafford, 1998, p. 185). In exploring the potential for performance assessments to be culturally responsive, authors explicitly considered the role of technology. Durán (1998) described how opportunities to play the role of a "web worker" who constructs web pages could provide assessment opportunities and also considered an extended "silver screen" project where students develop a television production using computational media. Boodoo (1998, p. 218) considered computer-adaptive performance assessments and recommended an emphasis on construct validity where the "construct is defined to include considerations of both the tasks and the test taker." Lee (1998) made recommendations for how performance assessments could be culturally relevant:

- integrating with curriculum
- drawing on funds of knowledge
- addressing a community-based, authentic need
- demanding complex skill and knowledge from multiple disciplines
- involving students in working together with those in- and outside their schools

Lee (1998) considered the above to be a cognitive rationale for culturally responsive performance assessments and provided a parallel section with a cultural rationale. This discussion

suggested incorporating movement and music, “verve” (high intensity), unique and genuine personal expression, orality, and affect. Although this special issue was published some time ago, a more recent review on, “Equity and Assessment: Moving towards Culturally Responsive Assessment” (Montenegro & Jankowski, 2017) continues the themes in Lee (1998). In particular, active research continues on the themes of performance assessment, construct validity that includes a student-centered point of view, drawing on funds of knowledge, and expecting complex skills from students.

Culturally-Relevant: Hits

It is premature to evaluate Beats Empire as a success with regard to cultural relevance; these reflections are therefore limited to observing *alignment* between the Beats Empire team’s intentions and field site experiences, and the call to action for culturally relevant assessments discussed above. In our designing and testing of Beats Empire, the team intentionally proceeded with learner (youth) culture in mind and with participatory design in school settings in New York City with high proportions of non-white students. The choice to focus on the music industry content and a game modality came out of this inclusive participatory design process because knowledge of music and skill in playing games are familiar and well-developed assets among students throughout New York City schools. Similarly, as recommended in the literature on culturally-relevant assessments, our playful assessment was designed to be a performance assessment by providing students with a challenge where multiple sources of knowledge could be brought to bear, and where complex skills could be demonstrated. For example, managing an iterative cycle of making decisions, examining the results, and using data to plan a more effective strategy. These particular constructs in the assessment were selected to align with computer science curriculum through the CS K-12 framework.

Further, although there is more to be done, the team's participatory design process intentionally engaged student and teacher voices. (See Chapters 2 & 8.) For this reason, specific features of the game—such as its geographic neighborhood map—were intended to directly suggest New York City. Similarly, the names of the artists in the game were chosen to suggest artists with whom students would be familiar. Students readily recognized the location and the connection to real artists that they listen to. Likewise, music genres represented in the game were those explicitly suggested by New York teens during participatory design sessions. This was one way our research found that students indeed bring cultural assets to how they want to approach the managerial role in the game and their assets give them ways to reason clearly about choices the game asks them to make (see Chapter 6). Students are not required to demonstrate what they know and can do by answering conventional test questions, but rather are placed in a performance context where the data they collect, visualize, and make inferences from leads to consequential choices in the game. In explaining their strategies and choices—and their connections to the data visualization and game outcomes—students and teachers can perform and discuss what they know and can do. It seems fair to say that the design of Beats Empire aligned to student assets and allowed students to more fully engage with the assessment opportunities in the game on the basis of that knowledge.

Culturally-Relevant: B-sides

When contrasting what we designed and observed with Beats Empire directly with frameworks of culturally responsive teaching and assessment, it is clear there is much more to be done. Chapter 6 observes that there could be inequitable consequences of choice in the game; students might choose a focus for their gameplay that is misaligned with the competencies that the game seeks to report on, and thus the game might mis-measure what they know and can do (see also

Chapter 5). As noted going back to Lee (1998), the assessment validity issue of construct irrelevant variance can be intertwined with students' culture and choices.

We also note that Beats Empire aligns best with the first of Ladson-Billings' aspects of culturally relevant teaching—academic success. The design is not yet intentional about how it could allow students to demonstrate their positive cultural identity (e.g. related to data analysis and computational thinking). As discussed in the section on authenticity, this is definitely possible in the game context; for example, we interviewed industry experts who deeply interconnected their positive personal identity with their daily work in analyzing music industry data. Plans for how to allow youth to experience and express this identity within and through the game are not yet in place for Beats Empire. Ladson-Billings' third aspect, to identify and critique social inequities, is also not an explicit design element in the game. For example, culturally responsive teaching should enable students to recognize and address power imbalances and racist behaviors at work. There is a history of white managers who exploit Black artists, and while the managerial format and studio setting of this game could support discussions of this history, it goes unaddressed in the game or instructional design (on this topic, done intentionally, there is also the possibility of addressing misogyny in gaming culture). Conversely, there are rich legacies of how Black managers and artists have promoted their art and grown their followers—such as Hip Hop's dissemination of mix-tapes—that have not been included as a part of the Beats Empire storyline. There are also real-world issues that are potentially interesting and important to Black youth that could be brought into the game—for example, issues relating to who collects and profits from student data, controls student data rights, and makes “recommendations” to students. These could be addressed within the music industry

context and would tie into additional CS Framework concepts, such as the Impacts of Computing concept.

If more curricular time were available, it would be plausible to build on what already exists in Beats Empire to explore additional issues. Could there be a game that encourages students to interview in-game characters to understand different roles that use data in the entertainment industry? How about one that explores how structures in the industry have used data to exploit, not just to benefit, artists? Could there be a game where students can be an apprentice to a successful Black artist management team and learn the strategies they take to combat racism in the industry?

Culturally-Relevant: Reflection

The Beats Empire game design aligns with a subset of characteristics that are called for in culturally-responsive pedagogy, but cannot yet be said to be culturally responsive or sustaining. The most we can say is that it shows promise as a *milieu* for further investigation. Culturally sustainable pedagogy promotes the use of “asset-based” approaches over conventional deficit-oriented teaching methods, in which non-mainstream languages, cultures, and identities are presented as barriers to student learning (Paris & Alim, 2017). Culturally relevant and sustaining pedagogies are not identical but they share the common goal of ensuring that students see themselves and their communities reflected and valued in the content taught in schools (Muñiz, 2019). A future version of Beats Empire (or another game with a similar entertainment industry context) could dig deeper into the assets in the Black or Latinx communities (for example) that have allowed for artistic success and for recognition and countering of racism. It could overcome a deficit-oriented perspective by presenting a context that is not only asset-based in terms of resonance with youth culture, but is also asset-based in terms of revealing more about people of

color and women who are thriving in music industry jobs that involve data—and as discussed in prior sections of this chapter, incorporating features where students are cognitively apprenticed to in-game characters who are positioned as successful in their job, confident in their identity and able to recognize and confront racism in their industry.

DISCUSSION

This chapter has reflected on the hits and “b-sides” of Beats Empire across three important concepts in the initial design brief: authenticity, playfulness, and cultural relevance. For each concept, we discussed an existing literature that developed a strong body of relevant theory. In reflecting on Beats Empire with that existing literature in view, we found both clear connections (“hits”) as well as aspects of each concept that were underdeveloped but potentially within reach of the overall music industry theme. We viewed these as “b-sides” that might be further developed in future iterations of research and development with Beats Empire or other games.

Looking at the hits, Beats Empire achieved a threshold of credibility relative to each concept. For example, with regard to authenticity, we noted that the game aligns to expectations of three important stakeholders. It fits at least two of the constructs in the CS-K12 framework—data analysis and interpretation—quite well, and has potential to expand to data collection and storage. It was authentic in the eyes of industry experts who serve artistic teams by designing tools for analyzing music industry data and deciding how to grow an artists’ popularity. It also was authentic to students as a game and as a representation of an industry they have experience with. Likewise, with regard to playfulness, we observed that Beats Empire was eminently playful—a characteristic that serious games sometimes fall short on. Finally, Beats Empire aligns to at least some called-for aspects for cultural relevant assessments.

Students had assets with regard to gameplay, geographic knowledge of New York City, and knowledge of artists and music that they were readily able to bring to bear. They were then placed in a performance assessment context that enabled them to demonstrate and talk about complex skills and knowledge.

Yet, looking across the b-sides for each concept, we see potential that is still unrealized. Throughout this reflection, we acknowledged that not enough time was available to deeply explore each of these potential areas; here we speculate on what could be accomplished if more time and resources were available.

Starting in reverse order with cultural relevance, we reflected that Beats Empire was more connected to academic success than positive cultural identity or ability to identify and address inequalities. The current game could not be said to be culturally sustaining as it is too short an experience, and it has not been made explicit how it would strengthen students' cultures. Yet, with more time and resources, it might be possible to achieve these desiderata in further efforts.

With regard to authenticity, we observed that greater authenticity might be achieved in several ways. For example, by connecting to the playlist concept, by having in-game mentors or further developing social role play around the game, and by developing the connection to rewarding careers that use computational and data skills. With regard to playfulness, we observed that expanding playfulness from the game to extended activities was a positive move.

What is especially striking, upon reflection, is the coherence among directions for improvement suggested by reviewing each of the three concepts. It seems that developing the connection to mentors and careers could make a future version of Beats Empire

more relevant to developing student's positive identity. Further, the playlist concept might also make the game relevant to issues of inequity and social justice as it intersects with the following questions. Although artists can produce songs using data, who controls the playlists that gate their success in developing followers? What barriers for talented women and/or artists of color get in the way of their pursuit of followers, money, and successful careers? The music industry is an approachable, understandable context in which students could explore possible career identities as well as inequities in an industry they are passionate about. If there was more time available for students to explore projects related to Beats Empire, we could imagine small group role playing games or activities that relate specifically to looking at how students' self-identity is growing. We could also imagine projects in which students explore historical inequities in the music industry and consider what kinds of data they could collect and analyze to look at those inequities today.

Thus, we see potential for powerful synergies to arise by digging deeper into the authenticity, playfulness, and culturally-relevant/culturally-sustaining constructs—especially as these different theoretical frameworks cross-connect. Narrowly, Beats Empire is a game in the management genre within a context that is both familiar and exciting to students. It is a proof of concept that authentic, playful, culturally responsive assessment is within reach. Yet we've also seen that by leveraging a deeper sense of authenticity (e.g. more authentic to reality in the music industry), a playful way to explore student connection to positive future identities may be possible. We've also seen that by expanding the playfulness of the game to activities that transcend the game—whether teacher-led, small group role playing activities, or activities with a mentor—it might be possible to create a playful, low-risk space for students to explore roles they could play with data beyond the artist management role. We've also observed that the music industry has abundant histories of

inequity that could be a source of authentic (and perhaps even playful) exploration. In conclusion, we argue that further attention to the cross-connections among authenticity, playfulness, and culturally responsive/culturally sustaining frameworks would yield a powerful design space for exploring the potential long-term contributions of playful assessments to the long-standing educational challenge of designing culturally relevant performance assessments.

CHAPTER 11.

AN INTERVIEW WITH GAME DESIGNER LUKE JAYAPALAN

SUMMARY

In this chapter we interview Luke Jayapalan, Senior Game Designer at Filament Games. In an interview between members of the research team and the lead of the developing team, we explore the development process, team dynamics, game features, and design challenges. We additionally compare our original concept of the game with the end result.

Matthew Berland (MB): *The biggest question we have for you – and something a lot of our chapters revolve around – is that we are particularly interested in the design tensions between assessment games and learning games. Could you just talk from your perspective about how you navigated that tension?*

Luke Jayapalan (LJ): So, from my perspective it kind of felt like the difference was partly just how much is it expected or acceptable that the student struggles a little bit to be successful in

the game environment. With a learning game, often the student struggling is seen as “We’re doing something wrong.” We have to very carefully scaffold every concept because we’re not presupposing that the players know any of the concepts. So part of what I think felt different here is that we still try to make sure that you’re understanding what you’re supposed to do, but we’re allowed to presuppose that either the student understands how to interpret and make predictions with the data that they’re looking at or they don’t. And if they don’t, that’s what the game is meant to uncover and to clarify why or in what areas they’re struggling. But it’s not specifically the game’s job to immediately resolve that issue or gate progress in the game, because they didn’t know how to do it.

MB: Yeah, that makes a lot of sense to me. I think there were also points where we disagreed – maybe disagreed is the wrong word...

LJ: Yeah, we had some ping pong of anxiety back and forth over how much a particular usability issue that we felt we had was going to interfere with assessment.

MB: Do you have any examples offhand? Because I think that’s really evocative.

LJ: I know they all had to do with looking at the data when graphed... Probably one difference that might have come up was with those predictive tools. There was a predictive tool for whether a song was going to be the highest, the most popular, or if it was going to be trending upward. There were times when I remember saying things like, “If we do it this way, then maybe students won’t be sure how to read this or make a successful prediction.” The two of you would say something like, “Maybe that’s okay.” It would just blow my mind, a little bit, because I think that is not usually the way a learning game works. I think it makes sense in the assessment context but....

MB: Right! So, you were saying, in the graphs, we’d come across a

situation where you would say something like, “But what if they don’t know how to interpret this graph, they won’t be able to progress.” And we would say, “That’s maybe a good thing!” But your natural instinct as a learning game designer is that this is inherently a bad thing and that we should scaffold them so they would not get stuck.

LJ: Yeah, if I go to a play test and I’m watching students use a learning game, then I worry that anything that they’re struggling with is my fault. We understood that it’s an assessment game and that it’s different, but I think we probably just had a different sense of when we could absolve ourselves of that responsibility.

MB: *Right! How do we know that the struggle was a conceptual struggle and not, say, a user interface struggle?*

LJ: Yeah! For example, we struggled with how to present that the game begins on a certain date, but in order to make predictions, you needed to have some past data. There was this struggle to come up with how we label data in the past. We contemplated having actual dates – the game begins on January 1 or whatever. At some point it was going to be negative numbers – “this is data from a week ago” – that’s all very confusing. Maybe we’ll just leave all the past data unlabeled entirely and it’s just blank? That was an example where if you’re used to seeing data plotted on a graph, you might understand that there is an ocean of past data.

The data collection frequency, that was another one that was tough. What components of the data should students be able to collect or not collect? We went through a lot of different rounds of revision. There were some conversations of whether or not we would have distractors – garbage data. I remember that getting talked about because, from an assessment perspective, you could say, “If they turn this on, it’s bad.” On the other hand, it is misleading... A game environment usually does not have anything that has literally no purpose! Instead, we tried to focus

more on: “This data is potentially useful, but in different situations. Can you use it in the appropriate situation?”

Nathan Holbert (NH): *I think you know from your experience with us that our team has varied opinions and diverse expertise, including a fair amount of prior experience with game design. I'm curious how your process changes when you're partnered with people who offer a lot of input throughout the design process?.*

LJ: Yeah, having game design experience makes everything more fruitful and straightforward because it's a debate grounded in a shared way of talking. Actually, a great example now that I think about it: the way we eventually came up with the trend-predicting tool. We knew that we wanted to have some way for the students to label the data to show that they're *understanding* the data and not just looking at it and making a gameplay decision. We really struggled with that. We had a mission structure at one point which was much more prescriptive and the student was going to have to do the mission; and I actually really appreciated that Nathan, was really pushing us to not have it be that kind of model where it's like, “Here's this prescriptive thing you have to do and at the end we're going to hide a little quiz in there to check to see if you understand what you're doing.” So having some shared philosophies really helped there.

MB: *That was not the answer that I was expecting. In part, because I know we were at least a little annoying to you.*

LJ: I think what you're referring to is, at least for Filament because we're a for-hire company, there's a certain amount of interpretation of what the clients' needs are. And I think a lot of times when we would have the back and forth, because our understanding, our interpretation, of your goals would say, “Wait if these are your goals, then don't we have to do it this way?” And sometimes you would come back with, “Yes, those are our goals, but we think we can do it this other way.” And we would

just not be used to the other way of doing it because there just aren't a lot of assessment based games out there. There are a lot of assessment based candy coated quizzes and there are a lot of games that aren't assessment based, even if they are learning games. But trying to do something that actually felt like it leverages all the systems that a rich learning game would for assessment purposes is not actually that common. So sometimes we would react with, "What do you want us to do?!" because you have to come up with a little more innovative solution than we're used to. When we hear, "We're going to need to know, 'Why did the student make the choice?'" we think, "We've got to put a moment in here where we ask you, and then there's a set of options, and then you pick the option that you used." But the prediction mechanic, where students label whether they chose a song's attribute because they felt it was the most popular, or trending up – and the player does this to maximize the earnings of their music studio, was really nice! And in that sense, it was more of a full mastery of the system.

***MB:** In academia, we're interested in generating new knowledge and that's very different from a lot of industry. A lot of industry is about finding success and creating successful products and that success is often measured in, "Did they sell? Is the client happy? Will the client return?" Which are more manageable, or at least quantifiable as goals. And I guess one of my questions is about you, Luke, which is when you undertake a project like this, what are some of your personal goals?*

LJ: Well, I always love it if I can see a kid playing – if they make a comment that something about that experience stuck with them or gave them a different perspective on something. Sometimes a student said something that's heartwarming or funny and that gives me a lot of satisfaction. As far as how I tell whether or not I feel like I did a good job, that is really tough to articulate because I think creative people are just tuned to like have a lot of feelings: you just look at it and you're like, "When I look at it, do I feel happy with where it is at right now? Or do I not?"

But the things I'm really proud of with this project is, I think that it was one of the most outside the norm attempts to do something really new with the merging of game interactivity and the educational space. It's trying to do something very innovative with assessment. That's something I really loved – I love the feel of the game. In terms of being a music studio: the idea of growing your set of artists, and all the bands having these cool names. And Josh Bartels wrote some really awesome music for the game! There are all these little clips that are not only different music genres but emulating different moods, so there's a sad pop song and the angry pop song and a happy song; that was fun!

I don't have any anxiety at all over if junior high students in the New York City school district will find this space an appealing gameplay space. And this is like perfect because there are data scientists in the music industry that do exactly this job! So the identity aspect of it is really strong – it might expose a kid to a career that they never thought was possible or existed. Those are the things I loved.

***NH:** I'll tell you my favorite moment, the one that really hit me, was these kids are sitting in groups of four or five and there were these two girls that each have their own laptop, but they were sitting next to each other. They each had their headphones in but every time one of them would be about ready to release a song she'd nudge her friend, they'd share one set of headphones, and then listen to the new song as it was released. And then they'd look at each other and both grin, nodd, and bob their head while they listened. They were so invested in the songs that they were recording in the game that they needed to show their friends!*

***MB:** My moment is, I took it to my son's school in Boston and one of the kids seemed pretty into it, he sort of just went off and started playing and he was playing during the breaks. For the next two weeks I was getting texts – I guess my son had given him my cell phone – about how*

much money he'd made and how many gold records: "Hey Matthew, I got all the top 10!"

NH: That's awesome! So I want to go back a little bit to something that we started talking about before. We mentioned the fact that there was a pretty diverse team of people on the academic side that we're working with. It was also a pretty big team of people. I don't know how common that is. How did that impact the kind of work that you all did?

LJ: The only real difference with having multiple points of contact for my client is that in this case all of the stakeholders on Beats Empire were able to advocate for different parts of the experience. And in that sense, it's quite similar to how our team at Filament works. We have a game designer, UX designer, engineer, etc. And usually, if we have a question it would naturally go to like one person or the other based on their discipline. But there are times when you have to talk to each other and there'd be times when the PFACS stakeholders have to talk to each other. It's just that there are more PFACS conversations that have to happen to get all the goals prioritized and sorted.

MB: So the more typical situation, a corporation or organization might contract a project that you are the designer on, and you would have contact with one stakeholder from that organization?

LJ: I don't know if there is a typical, to be honest. We try to have the client identify a feedback manager and so that the feedback manager is there to be the end point of the funnel. That sometimes means that – depending on how the client we're working with wants to organize things – sometimes that feedback manager is the only person that we talk to; a good number of clients, the weekly meetings consistent of two to four people; it's not unusual that stakeholders drop in at particular times, so you might have a subject matter expert who's not available regularly but they'll come in to review at some point.

So I think it wasn't tremendously different. I think the main thing that was different is that all the stakeholders were coming from an academic focus. And I think that in the academic environment, maybe, rigorous thinking-through and the debating things is more common and healthy. Being part of those debates gave us a lot of transparency into what were you guys were sure about and what you guys were not-so-sure about, because we can see those discussions happening live.

MB: That is something academics are good at – though I don't know if that made for efficient work on Beats Empire.

LJ: There were times when I would intentionally step back and wait and see, and then eventually Nathan would make some kind of call – and that's probably the only difference. I think, as far as the Filament side, it made it easy to talk directly with Daisy, for example, about, "We are having a hard time putting the storage objective in there." "What kind of variables do we need to store?" And that piece sped up communication. We built a lot of game in the amount of time that we had. With some clients, they have different needs, it may have certain corporate approval processes that you have to go through – that takes longer. It means that you're going to make a smaller amount of content in the same amount of time. Here I think we were able to move pretty fast because everybody that was a stakeholder was in there to raise their vantage point immediately and have it talked about and acted upon.

MB: So it seems that the timeline didn't stand out to you as particularly slow or us needing a lot of time to discuss things?

LJ: I don't think so. Often in the design phase – somewhere along the way – there are some design goals that nobody has quite figured out how to address. In Beats Empire, one of those was storage and the other was on how much exploring the data was critical to automate. Like, "Should the player have to make

some kind of automated flagging system?” That would have them essentially scripting, to have the game tell them, “Once this genre is popular, you’ll get a notification because you’ve set it up for yourself in the past.” Sometimes I wish that we would take even longer so that when we identify things like storage or automation – designs that are less clear – we have more time to hash it out. There’s a book that I read by Sabrina Culyba (2018) where she talks about how this is a pretty common problem for transformational games. If you know you’re going to hit all these goals that are very ambitious and atypical for games, then just having a lot of time in the design phase would be the way to go. But then it’s always paired against this problem that the budgets for transformational games are usually small; so that’s always the balancing act. So I think that the design phase always feels too short to me. That’s not unique to this project, that’s like every game: something comes up along the way where you’re just like, “Oh no, I wish I had more time to think about how to incorporate this!” And then you just run out.

NH: We want to ask you about trying to design a game that’s culturally relevant to students. I know our team thought quite a lot about whether or not the design that we all were working on – the art you were coming up with, the mechanics that everybody was coming up with – met the expectations and interests of these New York City middle school students. One obvious example: at the very beginning, pushing your team to think about the musicians beyond the five piece band – you could have solo artists, you could have a DJ, etc. Can you say a little bit about how that desire to think about making this game culturally relevant in New York City impacted your design process?

LJ: Well, we really loved getting the personas from you. A lot of transformational games in the US start from overly broad assumptions around who might benefit from it. So in some ways, it was nice that the Beats Empire project was naturally targeted. If you tell a very general story, and you make it as generic as possible, that doesn’t actually make it more relatable. The way

to make something relatable is to make it more specific, and if you include details from your own life or from other people's experience – a specific person's experience, then people pull out the parts that resonate with them, like "Oh yeah, I've had something like that!" even if it wasn't literally that same experience. I think, similarly, it was better to embrace, "This is going to feel like a city. It's going to have a certain kind of fashion sense, it's going to have a map." In your mind you could even extrapolate that to be like, "Oh, it seems like they're really into rock over here." It was the right amount of abstraction, where there wasn't any real attempt about specific identity, specific areas, and getting into stereotypical stuff. I think it's a lot better than if we had said, "Well you know, maybe there'll be some students that want to have folk music, or country music, or whatever..." and try to support every genre of music possible under the sun. All those things have a place, but I think that a more distinctive personality really helped as far as the student perspective.

NH: We shared some of the concept art your team was creating with local students early in the process to get their thoughts and feedback and they loved it!

LJ: I remember there was some focus testing done up front that led to those personas that we received, and another round after we put together some of the design materials. And maybe the reason why I don't remember a ton after that is because, we didn't really have to make a lot of pivots. Nobody came back and was like, "Oh.. God, you know, I really... that just didn't appeal to me at all." They might have different feels about what kind of music they listen to – and that gets into that specific versus general thing. Even if it wasn't all the music that you would love most, in some ways that's fine because it just created this world, and it felt like a coherent enough world within itself that you run with it. I think there's something for everybody.

***MB:** Excellent! Thanks again, Luke. This was super helpful! I think people are going to really enjoy reading your thoughts, and as always, it was a joy to work with you.*

CHAPTER 12.

CONCLUDING THOUGHTS

SUMMARY

In the “oral history” format, members of the research team reflect on the challenges and successes of creating and implementing the Beats Empire project, and, as a group, map out new research areas and design possibilities that we hope emerge from this work.

Throughout this book we have tried to reveal the motivations behind the development and study of Beats Empire, and provide specific examples and details about our process in order to offer a useful case study of assessment game design for audiences ranging from game designers, to curriculum developers, learning scientists, and teachers. Chapters in this book describe the research and co-design work done to understand the learning context and need, descriptions of core game design features and their theoretical underpinnings, details about the assessment design and the analysis of log data generated from gameplay, an accounting of the dashboard and bridging activity designs, and a reflection on our successes and missteps. The three-year

process was exciting and fun, while also occasionally frustrating and difficult.

For this last chapter, four of the team leads gathered together to offer some final thoughts about this extensive project. In this edited conversation Matthew Berland, Betsy DiSalvo, Daisy Rutstein, and Nathan Holbert reflect on the early challenges to communicate our vision to the game designers at Filament Games, discuss the tensions and opportunities in creating a playful constructionist assessment game, and offer a few lessons learned for curriculum, assessment, and learning design professionals.

BUILDING A CONSTRUCTIONIST FORMATIVE ASSESSMENT GAME

Betsy: So, I think initially we actually really struggled to come up with how to make an assessment game that was fun and included a constructionist approach to learning. I feel like we had, I don't know, four meetings with Matthew, Nathan, myself, and Jeremy [Roschelle, Digital Promise, who could not be present for this conversation]. We just kind of kept going around in circles and I thought that was frustrating.

Matthew: So the unexpected problem was that we had a theoretical position, but we didn't have a plan of attack that enacted that theoretical position?

Betsy: I don't think that's incorrect. I think that we had a theoretical position, but I think that I had an interpretation of that theoretical position that did not map to Matthew and Jeremy's position. I think Nathan and I were on the same page. That was my perception.

Matthew: Oh, so the difficulty was that I was part of the group. That seems fair. [Laughter]

Betsy: No, it's not bad at all! It just took us a while, because we come from different backgrounds, to figure out, "Okay, here's actually our take on this assessment." We actually didn't have any struggles recognizing, "Oh, they need assessment!" That's kind of the biggest thing we saw CS4All facing in New York (discussed in Holbert, Disalvo, & Berland, 2020). But I think our first reaction was, "God, we don't want to do assessment because we don't find it that exciting." So we thought, "How can we make assessment exciting? How can we make it fun?" And I think that Nathan and Matthew were like, "Let's make it constructionist!" And that started the ball rolling in terms of ideas.

Nathan: I think that's right and I remember distinctly – even in the earliest days, even before we probably settled on music – there was a lot of debate about what an assessment game could look like. Jeremy suggested a series of mini games. And the rest of us were like, "We don't want these disconnected experiences. We want players to engage in a longer process of building something interesting or meaningful."

Daisy: For me, there was a lot of coming to terms with what it meant to create a game to be a formative assessment. For example, I was sort of surprised at how hard it was to figure out the balance between how much we could and could not guide students in the game. The idea that we couldn't let students fail was new to me! I remember we had a conversation where players would make choices that would generate the data and someone from Filament asked, "What if the player made choices that meant there was nothing in the graph?" I was like "Well, then there's nothing in the graph! That's good!" And there was a lot of pushback on that! For a game, we didn't want to block them, or for players to feel stuck, we wanted them to still be able to progress in the game. This meant we wouldn't always get the clear data of, "Yes, they can do it, no they can't." Instead, the data we'd get was fluid, there's more room. That was something I know I grappled with a lot.

ALIGNING THE VISION OF GAME DEVELOPERS AND GAME RESEARCHERS

Nathan: It's worth noting, there was occasionally conflict between our team and Filament Games, the developer. While we had our particular ideas about what a playful and constructionist formative assessment could be, Filament also had their perspective and their priorities. Unfortunately our perspectives didn't always align.

Betsy: The first pitch that they gave us for what they were going to build – I remember being actually upset. They basically pitched “Battle of the Bands” and I was like “I think we're fucked.” That's what I really thought. Because it was so not what we had put forward to them, and I also didn't feel like anybody else was upset.

Nathan: I'm with you. I remember really struggling with that first pitch from Filament as well. I had come up with a lot of ideas that I was excited about for the game which we used to describe the project to Filament. At the core of this was that we would build, essentially, like, an in-game querying language. The players would write pseudo code to query the data then the game would generate the representations they'd use to make decisions. And those decisions could be for song recording, for the social media team, etc.

And from that idea, they came back to us with their battle of the bands game pitch, it was like, “Wait, what? This isn't – this is nothing like what we asked for!” I wasn't angry, I don't think, or as horrified as you, Betsy, but I was certainly like, “Okay there's a huge gulf between what we were thinking and what they're thinking, and how in the world are we going to bridge this?”

Matthew: Yeah I didn't like the first version. But my feeling with game design is always like, the first 50 versions are total failures,

and then you sort of eke out something. So I remember thinking, “No. But we can start there, that’s fine.”

Betsy: That’s the right attitude – But I didn’t know Filament as well and I was concerned. I feel like they weren’t thinking at all about what it means to build a game for the huge diversity of kids in New York City. I thought maybe we should bring them to New York to talk to kids there.

Nathan: You were quite honest about being frustrated.

Matthew: Yeah, Betsy was very clear and that was very helpful! Betsy, you were like “This is why this is not acceptable and here’s where we need to go.” And they were like “okay.” And one part of the solution was a change in the staffing. Then Luke, the game director, and Megan, the visual director, came on board. And then the leadership of the project became entirely people of color which also changed the project for the better – better reflected the students..

Betsy: Yeah [Filament] responded amazingly, and the whole process totally changed.

BUILDING AN EFFECTIVE TEAM

Nathan: I was really quite happy with and learned a lot from the ways in which this team worked together. We had a really interesting collection of people with different expertise! Early on we all kind of agreed to be equally responsible for all of the different parts of the project, as opposed to a more typical divide and conquering model. That meant we all probably invested a lot more time in this project than we might have otherwise. But I think it also meant that I got to see a lot of the work that you were doing, Daisy, with the assessment stuff, and so I was able to learn a lot from you! We were also able to have a huge amount of input on the particulars of the game, as it was being developed by Filament. So we didn’t just design it and then leave, we were

able to impact the development phase. The participatory design work that we did at the beginning of the project – and all of us having some input on those experiences and reflecting on the implications of it. Just throughout the entire project it felt like we had everyone’s brains actively working on all of the many different parts of what we were creating! That led to all sorts of tensions certainly, and led to some great debates and good natured arguments. But I also think it led to a pretty cool design and a pretty cool project! I haven’t had another team, especially one this big, work quite so well together. So it’ll be interesting to sort of see if this model can be replicated in other places. Or maybe we were just fortunate to all be delightful people, which we all are, of course.

Betsy: It was a surprisingly collaborative group for this size, because you’re right when I normally work on projects that have this many people, it’s like everybody goes off in their little silo and does their thing. And some of it, I think, had to do with the fact that we were meeting weekly! We were meeting together, I was meeting with my own team, or with other people related to the project...

Nathan: We definitely put a lot of hours into this project for sure.

Daisy: I agree, we have really good conversations and we had lots of tensions, but we were able to work through them, and talk about them, and be open about everything. And so I think that led to being pretty productive.

Nathan: We did have tensions! But they were productive and generative tensions. We didn’t resolve every tension exactly like each of us would have hoped probably, but I felt like these tensions raised really interesting questions about assessment, about play, about design, about computer science, about data. We got into some pretty big issues – it was fun!

LESSONS FOR CURRICULUM AND EDUCATIONAL APP

DESIGN

Betsy: So we have a lot of potential audiences for this work. One audience might be people who are game designers who are asked to do educational games. I think that's one audience.

Nathan: What about curriculum developers that are starting to sort of dabble in – or being instructed to dabble in game based learning or gamification for things to be used in schools? What do we hope their takeaways are for Beats?

Betsy: I think two things I learned that are worthwhile is first, to work with real game designers, to design something that is of a high quality. And second, assessment can be gathered from game data, but it has to be delivered to teachers in a way that is accessible to them.

Nathan: I think a mistake we often make when we are trying to design games or other learning environments, is we think “What math do they need to learn?” And our response to that is to look at what math is currently being taught in schools. But really, what we should be thinking about is what is the particular idea or skill or practice that's valuable and useful beyond the classroom and how can we design a set of experiences around those real-world practices. And in our case with Beats, we tried to think about where do kids encounter data? Where are they excited to think about and reflect on data? And we tried to build a game in a music context that touches on some of the same kinds of practices and skills and ideas that a computer science classroom or a math classroom might be interested in supporting. But we try to think about it in a meaningful context that exists beyond a classroom.

Daisy: Yeah, I agree with that because I think one of the benefits of the game is being able to have students in situations that have that sort of a real-world equivalent. Players are not just displaying their knowledge – like “Do you know it or not?” –

they're getting the chance to use ideas in interesting situations. Games really offer a powerful opportunity to get at things that you can't get at on a traditional assessment.

Matthew: Right! I think an ultimate success might be that we add a tool to the arsenal of many of these companies building assessment apps; that they might think, "Oh well, this might help us get at these other angles, and we might not get all the data that we would love to get about that for a test or whatever, but it'll give us a better understanding of this or that in a way that feels connected to the students. We could throw a game designer and developer at that for a few weeks, see what happens."

Daisy: But one thing I hope readers take away from this book is just an appreciation for just how challenging this can be to get it right. And how there's just so many different aspects to the game that you need to consider. You have to find a balance between making sure that it's fun and engaging, but also useful to teachers. You have to provide relevant information, and also you need to make sure you're measuring something that's important... It all requires a lot of tradeoffs – and it's hard to do all of these in one game.

Matthew: Yeah, I think that's true, I definitely think that's one of the things that I hope people take away. But the flip side of that is showing the value in this approach, such that people might want to build on it and take it to new places. And then have an easier time, hopefully, than we had because they'll be able to build on some of the challenges that we saw.

Nathan: To put it a different way, I hope Beats Empire can be an example and also permission to try to think about assessment in a different way. People that don't know a lot about assessment – and I would include myself before this project in that category of people – get kind of locked into an assumption about what an assessment has to look like or what it's supposed to be. And I

think we pushed on that notion with Beats Empire. We showed that you can have an open-ended assessment, you could have an assessment that was playful, that was actually meaningful to learners, but still gave students and teachers some potentially important information about what they're learning and what they haven't yet captured.

Betsy: I think we're all thinking along the same lines, but I do think maybe I'm agreeing more with Daisy. One of the reasons why I think parts of this game were successful is because we had a variety of different disciplinary approaches and expertise on this team. We knew how to make a game that is appealing to kids, but we also thought a lot about the teachers. And because we had assessment expertise, we were able to think about how interesting game experiences would elicit assessment information. And the dashboard – the user interface for the dashboard is immensely important in my opinion! I came out of this process thinking you can't have a game for the classroom that doesn't have a dashboard. And I never really thought about educational games having a teacher dashboard before this.

So I don't know if it is that easy and everybody should try it.

Matthew: I'm agreeing with you and Daisy. But just saying that I hope that we made it easier for other people because it is hard. But I also think it's worthwhile.

I have the perspective of someone who has a child in middle school, and he spends a lot of his school day on things that maybe would have been considered games 25 years ago. Like quiz games and this kind of thing. But they're all very much the same, which is very surprising to me. The sort of theoretical gap between them, despite being made by different companies with pretty different perspectives on the world!

This morning my son and I had a funny conversation. I asked him about the difference between two game apps I saw him using

for school. And he goes, “Well, one of them sort of tries to teach you by giving you the problems. The other just sort of gives you the problems, but then has a fun way of telling you how you’re wrong.” And I said, “Which do you like better?” He goes, “They’re all too easy, and they can be boring. So I guess I like the one that lets you just do the problems, because if you can sort of manage not to get any of them wrong, you can make it go way faster.”

I think that there is a version of playful assessment which has taken hold in public middle schools across the United States at the very least – and I would be surprised if it wasn’t elsewhere as well – where it is just traditional school worksheets, but with the added feature of allowing you to fail. I do think that’s better than not allowing you to fail (by a ton), but it’s still not open in any way.

Nathan: This is a really nice point you make about sort of the overwhelming nature of gamified and digital homework that’s become super common in schools. And this was accelerated by all the remote schooling that was going on with the pandemic.

Betsy: Right. It definitely seems true that there are a lot of companies that are producing assessments that are gamified, but the game is about earning more points by doing better in the assessment. The assessment isn’t tied to the game mechanics in any meaningful way.

LESSONS FOR LEARNING GAME DESIGN

Matthew: On the flip side of that, what lessons do we have for learning game designers? When Luke asked us, “What is the difference between a learner game and an assessment game?” We didn’t have anything to give him! But now we do have an answer to that question.

Nathan: Right. I think there’s some pretty important overlaps between those two types of designs, but also some very

important distinctions that weren't very clear to me before we started this process.

Daisy: Basically, there's also a lot of rethinking from an assessment perspective of what it means to develop assessment and what are the rules. And maybe they're not as strict as you think they are. There's room to kind of widen what you consider "assessments".

Betsy: I tend to think all learning games should probably have some assessment built into them.

Daisy: I'd say yes! That's my biased perspective.

Nathan: I don't know... I'm not convinced that's true. But I'm not offended by that statement.

Betsy: I mean how about this? Maybe not all games should have assessment built into it so that the students can see it, but they should provide information about student performance for the teachers.

I think that games are an opportunity for formative feedback so the teachers can understand where students are. I mean, having games in classes is not anything new. We played games when I was a kid, before there were computers in the classroom, right? I think of the many things these games did, they did because it was a performance – the teacher could see how you were doing. They could see how the students were engaging with the material and what they actually understood. I think the problem with screen based games is that it's really hard for teachers to capture that.

And so we should think about how to make that performance visible, especially in screen-based games, when we're designing learning games for the classroom.

LESSONS FOR FORMATIVE ASSESSMENT DESIGN

Matthew: I remember many conversations that – I think we’re still having even as of this morning – about, “What is formative assessment?” But I also think there’s a question of, “How much does Beats Empire look like a formative assessment?” I looked again over the research on formative assessment in the last few days and Beats Empire, and this larger project around the game, really fits the definitions generally given for formative assessment, but doesn’t look like any of the work in that space. All the definitions are like, “Formative assessment is how you see where people are, and help the teaching get better, and help the classroom be more productive, and help students learn, etc.” And then the actual projects don’t really share that many features with what we did. Where’s the disconnect? I think that my biggest hope is that we could just insert ourselves into that conversation a little bit more and press people to open up what they count as formative assessment. Because I think that’ll have multiplicative effects down the line, if we can convince people this is a valid form of formative assessment.

Daisy: Yeah I do think that what we did in this game is very different from how formative assessment normally is presented. And even from game-based assessment! A lot of the game based assessments are little mini games and you can clearly see these are assessments. I mean you put a little character and fun stuff around it, but it’s really just asking kids math questions. You can broaden your idea of assessment and what it means to do a formative assessment. It doesn’t have to be this rigid, “Can you answer this one multiple choice question?”, and that kind of thing. This widens it and says there’s ways to capture this information, but it doesn’t have to feel *assessment-y*. Or feel like students know that, “Oh, there’s a right answer and I’m supposed to put down the right answer,” but that students are also allowed to have a variety of answers and ways to go about things.

Nathan: I think that's right, but I think in some ways it's not so surprising, in part because of the frictions that we encountered in doing this. To create something that's quite interesting and quite fun and does offer some useful formative assessment moments; it's actually pretty freaking difficult! And so there isn't a lot in this space, there's this lack of examples for people to look to.

Matthew: It is a sparse space. Like, of the formative assessments that I've seen, maker formative assessments, where they evaluate artifacts, is the one that looks most like Beats Empire. Or artistic formative assessments, where they evaluate intermediate representations and things. But this stuff doesn't show up in science or math classes as often.

Betsy: But there *are* many similar formative assessments that aren't educational technology. I think teachers do this without even being aware they're doing it! They come up with a clever game or activity not just to teach something but also just to give them a feel for where people are in an informal way. Good teachers are already doing this! When we talk to the teachers, they talk about how they have impromptu kinds of presentations and things like that, and they try to make them fun or playful, because that gives them a chance to do formative assessment. It was one of the ways that they use the most – more frequently than I had anticipated.

Before beginning this project I had no background in assessment, and, I'm not saying that I am an expert in it now, but I certainly have a much greater appreciation for it!

Daisy: I have more of an appreciation of the game design side, and in developing fun experiences, and the challenges that came with that. I'm still not completely sold on game-based assessment. I was actually really excited about game-based assessment before and now I'm like, "Well, there's a lot that needs to go into this to get it right!" And that's challenging. And so I'm

not 100% sure that we want all assessments to go that way. But I think they definitely have their place.

Nathan: I think you'd find few people – and definitely no one on this call – that would believe that, “Yes, all assessments should be game-based assessment.” They have a particular use, but they also have some weaknesses.

CONCLUDING THOUGHTS

The Beats Empire team came together to address the specific need to provide formative assessment tools for NYC teachers implementing new computer science curriculum in middle schools across the city. The resulting formative assessment game was meant to produce actionable information about student learning around data and data analysis while also being culturally relevant and enjoyable to students. We believe this project demonstrates that games are one way to make assessment engaging and meaningful for students. And in addition to the direct classroom applications of Beats Empire, we hope the game and the design process articulated here can be useful for game design, ed tech, and assessment companies looking to go deeper than simply adding gamified elements to digital homework. Bringing these domains together can result in new, more engaging experiences that give a summative picture of student learning, but also can be used as a formative assessment, informing teachers' actions in the classroom.

Audience awareness was very important: from understanding what aspects of the game would be meaningful to players to understanding how teachers could use gameplay in the classroom. An effective formative assessment game must not only collect relevant information of student understanding, it must also communicate this information in a straightforward and clear way to teachers, enabling them to make thoughtful decisions about how to respond to this information by

leveraging their knowledge about their students, and their expertise as classroom teachers. Giving teachers guides for how to use the game as a tool in their class also matters – this isn't a case of “build it and they will come” – teachers can't be expected to do all the work discovering and fitting Beat Empire, or any game, into their broader curricular goals.

Educational game design is very difficult – we had a wide range of expertise from content area experts, assessment experts, audience experts, and of course expert game designers. As the conversation above highlights, it was not only important to have this diverse expertise represented on the team, but also to have team members frequently communicating, pitching new ideas, arguing about key values, and finding consensus. Balance was key to making the game enjoyable and also useful for learning. So while working with game designers and learning experts can be a challenge, with the push and pull between what the learning or assessment goals are, what the audience will respond to, and the vision of the designers – it is necessary and worthwhile.

REFERENCES

- Abras, C., Maloney-Krichmar, D., & Preece, J. (2004). User-centered design. In W. S. Bainbridge (Ed.), *Berkshire encyclopedia of human-computer interaction* (pp. 445–456). Berkshire Publishing Group LLC.
- Ahn, J., Campos, F., Hays, M., & DiGiacomo, D. (2019). Designing in Context: Reaching beyond Usability in Learning Analytics Dashboard Design. *Journal of Learning Analytics*, 6(2), 70–85.
- Andrade, H. L., & Heritage, M. (2017). *Using formative assessment to enhance learning, achievement, and academic self-regulation*. Routledge.
- Asbell-Clarke, J., Rowe, E., Bardar, E., & Edwards, T. (2020). The Importance of Teacher Bridging in Game-Based Learning Classrooms: In M. Farber (Ed.), *Advances in Educational Technologies and Instructional Design* (pp. 211–239). IGI Global. <https://doi.org/10.4018/978-1-7998-2015-4.ch010>
- Atherton, P. (2018). *50 Ways to Use Technology Enhanced Learning in the Classroom: Practical Strategies for Teaching*. Learning Matters.
- Bangor, A., Kortum, P. T., & Miller, J. T. (2008). An Empirical Evaluation of the System Usability Scale. *International Journal of Human-Computer Interaction*, 24(6), 574–594. <https://doi.org/10.1080/10447310802205776>
- Barchas-Lichtenstein, J., Brucker, J. L., Voiklis, J., Thomas, U. G., Fraser, J., Shane-Simpson, C., & Fields, S. (2019). *Cultivating computational*

thinking in elementary & middle school Learners (Knology Publication #NSF.051.213.05). Knology.

Basu, S., Disalvo, B., Rutstein, D., Xu, Y., Roschelle, J., & Holbert, N. (2020). The Role of Evidence Centered Design and Participatory Design in a Playful Assessment for Computational Thinking About Data. *Proceedings of the 51st ACM Technical Symposium on Computer Science Education*, 985–991. <https://doi.org/10.1145/3328778.3366881>

Betts, J. R., & Grogger, J. (2003). The impact of grading standards on student achievement, educational attainment, and entry-level earnings. *Economics of Education Review*, 22(4), 343–352. [https://doi.org/10.1016/S0272-7757\(02\)00059-6](https://doi.org/10.1016/S0272-7757(02)00059-6)

Black, P., & Wiliam, D. (2009). Developing the theory of formative assessment. *Educational Assessment, Evaluation and Accountability(Formerly: Journal of Personnel Evaluation in Education)*, 21(1), 5–31. <https://doi.org/10.1007/s11092-008-9068-5>

Blikstein, P. (2013). Digital fabrication and ‘making’ in education: The democratization of invention. *FabLabs: Of Machines, Makers and Inventors*, 4(1), 1–21.

Bogost, I. (2008). *The rhetoric of video games*. MacArthur Foundation Digital Media and Learning Initiative.

Boodoo, G. M. (1998). Addressing Cultural Context in the Development of Performance-Based Assessments and Computer-Adaptive Testing: Preliminary Validity Considerations. *The Journal of Negro Education*, 67(3), 211–219. <https://doi.org/10.2307/2668190>

Brouns, F., Zorrilla Pantaleón, M. E., Esperanza Álvarez Saiz, E., Solana González, P., Cobo Ortega, Á., Rocío Rocha Blanco, E., Collantes Viaña, M., Rodríguez Hoyo, C., De Lima Silva, M., Marta-Lazo, C., Gabelas Barroso, J. A., Arranz, P., García, L., Silva, A., Sáez López, J. M., Ventura Expósito, P., Jordano de la Torre, M., Bohuschke, F.,

- & Viñuales, J. (2015). *ECO D2.5 Learning analytics requirements and metrics report*.
- Brown, J. S., Collins, A., & Duguid, P. (1989). Situated Cognition and the Culture of Learning. *Educational Researcher*, 18(1), 32–42. <https://doi.org/10.3102/0013189X018001032>
- Cocca, M., Hershkovitz, A., & Baker, R. S. J. d. (2009). The Impact of Off-task and Gaming Behaviors on Learning: Immediate or Aggregate? *Artificial Intelligence in Education*, 507–514. <https://doi.org/10.3233/978-1-60750-028-5-507>
- Cochran, K. F., DeRuitter, J. A., & King, R. A. (1993). Pedagogical Content Knowing: An Integrative Model for Teacher Preparation. *Journal of Teacher Education*, 44(4), 263–272. <https://doi.org/10.1177/0022487193044004004>
- Cole, M., & Consortium, D. L. (2006). *The Fifth Dimension: An After-School Program Built on Diversity*. Russell Sage Foundation.
- Consalvo, M. (2009). *Cheating: Gaining Advantage in Videogames*. MIT Press.
- Cooper, A., Reimann, R., Cronin, D., & Noessel, C. (2014). *About Face: The Essentials of Interaction Design*. John Wiley & Sons.
- Crooks, R. N. (2019). Times thirty: Access, maintenance, and justice. *Science, Technology, & Human Values*, 44(1), 118–142.
- Csikszentmihalyi, M. (1990). *Flow: The psychology of optimal experience*. Harper & Row.
- Culyba, S. (2018). *The Transformational Framework: A Process Tool for the Development of Transformational Games*. Carnegie Mellon University: ETC Press. <https://doi.org/10.1184/R1/7130594.v1>
- DesPortes, K., Spells, M., & DiSalvo, B. (2016). The MoveLab: Developing Congruence Between Students' Self-Concepts and Computing. *Proceedings of the 47th ACM Technical Symposium on*

Computing Science Education, 267–272. <https://doi.org/10.1145/2839509.2844586>

DiSalvo, B., & DesPortes, K. (2017). Participatory design for value-driven learning. In *Participatory Design for Learning*(pp. 175–188). Routledge.

DiSalvo, B., Yip, J., Bonsignore, E., & DiSalvo, C. (Eds.). (2017). *Participatory Design for Learning: Perspectives from Practice and Research*. Routledge. <https://doi.org/10.4324/9781315630830>

Duran, R. P. (1998). Learning and Technology: Implications for Culturally Responsive Instructional Activity and Models of Achievement. *The Journal of Negro Education*, 67(3), 220–227. <https://doi.org/10.2307/2668191>

Ehn, P. (2008). Participation in design things. *Proceedings of the Participatory Design Conference*, 92–101.

Electronic Arts Inc. (2016). *SimCity*. <https://www.ea.com/games/simcity>

Elkind, D. (2008). The Power of Play: Learning What Comes Naturally. *American Journal of Play*, 1(1), 1–6.

Ericsson, K. A., & Simon, H. A. (1984). *Protocol analysis: Verbal reports as data*. MIT Press.

Frauenberger, C., Good, J., Fitzpatrick, G., & Iversen, O. S. (2015). In pursuit of rigour and accountability in participatory design. *International Journal of Human-Computer Studies*, 74, 93–106.

Frederiksen, N., Mislevy, R. J., & Bejar, I. I. (1993). *Test theory for a new generation of tests*. Erlbaum Hillsdale, NJ.

Freeman, J., Magerko, B., McKlin, T., Reilly, M., Permar, J., Summers, C., & Fruchter, E. (2014). Engaging underrepresented groups in high school introductory computing through computational remixing

with EarSketch. *Proceedings of the 45th ACM Technical Symposium on Computer Science Education*, 85–90.

Galitz, W. O. (2007). *The Essential Guide to User Interface Design: An Introduction to GUI Design Principles and Techniques*. John Wiley & Sons.

Gay, G. (2018). *Culturally Responsive Teaching: Theory, Research, and Practice* (3rd ed.). Teachers College Press.

Gee, J. P. (2007). *Good Video Games + Good Learning: Collected Essays on Video Games, Learning, and Literacy*. Peter Lang.

Gee, J. P. (2014). *What Video Games Have to Teach Us About Learning and Literacy*. Macmillan.

Greeno, J. (2005). Learning in activity. In R. K. Sawyer (Ed.), *The Cambridge Handbook of the Learning Sciences*. Cambridge University Press.

Grover, S. (2021). Toward A Framework for Formative Assessment of Conceptual Learning in K-12 Computer Science Classrooms. *Proceedings of the 52nd ACM Technical Symposium on Computer Science Education*, 31–37. <https://doi.org/10.1145/3408877.3432460>

Habgood, M. J., & Ainsworth, S. E. (2011). Motivating children to learn effectively: Exploring the value of intrinsic integration in educational games. *The Journal of the Learning Sciences*, 20(2), 169–206.

Hansen, N. B., Dindler, C., Halskov, K., Iversen, O. S., Bossen, C., Basballe, D. A., & Schouten, B. (2019). How participatory design works: Mechanisms and effects. *Proceedings of the 31st Australian Conference on Human-Computer-Interaction*, 30–41.

Haskell, R. E. (2000). *Transfer of learning: Cognition and instruction*. Elsevier.

Herrington, J., Oliver, R., & Reeves, T. C. (2003). Patterns of engagement

- in authentic online learning environments. *Australasian Journal of Educational Technology*, 19(1). <https://doi.org/10.14742/ajet.1701>
- Herrington, J., Parker, J., & Boase-Jelinek, D. (2014). Connected authentic learning: Reflection and intentional learning. *Australian Journal of Education*, 58(1), 23–35. <https://doi.org/10.1177/0004944113517830>
- Holbert, N., Berland, M., & Kafai, Y. B. (2020). *Designing constructionist futures: The art, theory, and practice of learning designs*. MIT Press.
- Holbert, N., Disalvo, B., & Berland, M. (2020). The Rollout of Computer Science Education to Every Student in New York City: A Socio-Contextual Social Network Analysis. *Teachers College Record*, 122(11), 1–30. <https://doi.org/10.1177/016146812012201106>
- Holbert, N., & Wilensky, U. (2014). Constructible authentic representations: Designing video games that enable players to utilize knowledge developed in-game to reason about science. *Technology, Knowledge and Learning*, 19(1), 53–79. <https://doi.org/10.1007/s10758-014-9214-8>
- Holbert, N., & Wilensky, U. (2019). Designing educational video games to be objects-to-think-with. *Journal of the Learning Sciences*, 28(1), 32–72. <https://doi.org/10.1080/10508406.2018.1487302>
- Hood, S. (1998). Introduction and Overview: Assessment in the Context of Culture and Pedagogy: A Collaborative Effort, a Meaningful Goal. *The Journal of Negro Education*, 67(3), 184. <https://doi.org/10.2307/2668187>
- Hunicke, R., LeBlanc, M., & Zubek, R. (2004). MDA: A formal approach to game design and game research. *Proceedings of the AAAI Workshop on Challenges in Game AI*, 4(1), 1722.
- Jankowski, N. A. (2020). Assessment during a Crisis: Responding to a Global Pandemic. *National Institute for Learning Outcomes Assessment*.

- Jivet, I., Scheffel, M., Drachsler, H., & Specht, M. (2017). Awareness Is Not Enough: Pitfalls of Learning Analytics Dashboards in the Educational Practice. In É. Lavoué, H. Drachsler, K. Verbert, J. Broisin, & M. Pérez-Sanagustín (Eds.), *Data Driven Approaches in Digital Education* (pp. 82–96). Springer International Publishing. https://doi.org/10.1007/978-3-319-66610-5_7
- K–12 Computer Science Framework*. (2016). <http://k12cs.org>
- Kim, Y. J., Almond, R. G., & Shute, V. J. (2016). Applying Evidence-Centered Design for the Development of Game-Based Assessments in Physics Playground. *International Journal of Testing*, *16*(2), 142–163. <https://doi.org/10.1080/15305058.2015.1108322>
- Kim, Y. J., & Miklasz, K. (2021). What We Learned: From Games to Make Assessment Playful. In *Teaching in the Game-Based Classroom* (pp. 151–161). Eye on Education.
- Kim, Y. J., Murai, Y., & Chang, S. (2021). Implementation of embedded assessment in maker classrooms: Challenges and opportunities. *Information and Learning Sciences*.
- Ladson-Billings, G. (1995a). But that's just good teaching! The case for culturally relevant pedagogy. *Theory Into Practice*, *34*(3), 159–165. <https://doi.org/10.1080/00405849509543675>
- Ladson-Billings, G. (1995b). Toward a Theory of Culturally Relevant Pedagogy. *American Educational Research Journal*, *32*(3), 465–491. <https://doi.org/10.3102/00028312032003465>
- Lee, C. D. (1998). Culturally Responsive Pedagogy and Performance-Based Assessment. *The Journal of Negro Education*, *67*(3), 268. <https://doi.org/10.2307/2668195>
- Lin, Q., Yin, Y., Tang, X., Hadad, R., & Zhai, X. (2020). Assessing learning in technology-rich maker activities: A systematic review of empirical research. *Computers & Education*, *157*, 103944.

- Lindblad, S., Pettersson, D., & Popkewitz, T. S. (2018). Education by the numbers and the making of society. *New York: Routledge*, 20, 9781315100432.
- Lui, D., Walker, J. T., Hanna, S., Kafai, Y. B., Fields, D., & Jayathirtha, G. (2020). Communicating computational concepts and practices within high school students' portfolios of making electronic textiles. *Interactive Learning Environments*, 28(3), 284–301.
- Matcha, W., Uzir, N. A., Gašević, D., & Pardo, A. (2020). A Systematic Review of Empirical Studies on Learning Analytics Dashboards: A Self-Regulated Learning Perspective. *IEEE Transactions on Learning Technologies*, 13(2), 226–245. <https://doi.org/10.1109/TLT.2019.2916802>
- McLelland, N. (2017). *Teaching and learning foreign languages: A history of language education, assessment and policy in Britain*. Routledge.
- Means, B., & Olson, K. (1994). The Link between Technology and Authentic Learning. *Educational Leadership*, 51(7), 15–18.
- Mislevy, R. J., & Haertel, G. D. (2006). Implications of Evidence-Centered Design for Educational Testing. *Educational Measurement: Issues and Practice*, 25(4), 6–20. <https://doi.org/10.1111/j.1745-3992.2006.00075.x>
- Mislevy, R. J., Oranje, A., Bauer, M. I., von Davier, A. A., & Hao, J. (2014). *Psychometric considerations in game-based assessment*. GlassLabGames.
- Molenaar, I., & Knoop-van Campen, C. A. N. (2019). How Teachers Make Dashboard Information Actionable. *IEEE Transactions on Learning Technologies*, 12(3), 347–355. <https://doi.org/10.1109/TLT.2018.2851585>
- Montenegro, E., & Jankowski, N. A. (2017). Bringing Equity into the Heart of Assessment. *Assessment Update*, 29(6), 10–11. <https://doi.org/10.1002/au.30117>

- Muñiz, J. (2019). *Culturally Responsive Teaching: A 50-State Survey of Teaching Standards*. New America. <https://eric.ed.gov/?id=ED594599>
- National Research Council. (2001). *Knowing What Students Know: The Science and Design of Educational Assessment*. <https://doi.org/10.17226/10019>
- Ndemic Creations. (n.d.). *Plague Inc*. Retrieved November 8, 2022, from <https://www.ndemiccreations.com/en/22-plague-inc>
- Newman, D., Griffin, P., & Cole, M. (1989). *The Construction Zone: Working for Cognitive Change in School*. Cambridge University Press.
- Papert, S. (1980). *Mindstorms: Children, computers, and powerful ideas*. Basic Books, Inc.
- Paris, D. (2012). Culturally sustaining pedagogy: A needed change in stance, terminology, and practice. *Educational Researcher*, 41(3), 93–97.
- Paris, D., & Alim, H. S. (2017). *Culturally Sustaining Pedagogies: Teaching and Learning for Justice in a Changing World*. Teachers College Press.
- Pellegrino, J. W. (2014). Assessment as a positive influence on 21st century teaching and learning: A systems approach to progress. *Psicología Educativa*, 20(2), 65–77.
- Pellicone, A., Holbert, N., DiSalvo, B., Kumar, V., & Berland, M. (2019). Who Played the Game Correctly? In J. H. Kalir & D. Filipiak (Eds.), *Proceedings of the 2019 Connected Learning Summit*. Carnegie Mellon University: ETC Press.
- Plass, J. L., Homer, B. D., & Kinzer, C. K. (2014). *Playful learning: An integrated design framework* [White Paper].
- Plass, J. L., Homer, B. D., & Kinzer, C. K. (2015). Foundations of game-based learning. *Educational Psychologist*, 50(4), 258–283.

- Prensky, M. (2006). *Don't bother me, Mom, I'm learning!: How computer and video games are preparing your kids for 21st century success and how you can help!* Paragon house St. Paul, MN.
- Pruitt, J., & Grudin, J. (2003). Personas: Practice and theory. *Proceedings of the 2003 Conference on Designing for User Experiences*, 1–15.
- RCTO Productions, LLC. (n.d.). *RollerCoaster Tycoon*. Retrieved November 8, 2022, from <https://www.rollercoastertycoon.com/>
- Roschelle, J. (2019, March 22). Introduce Computational Thinking with the Music Industry. *Digital Promise*. <https://digitalpromise.org/2019/03/22/introduce-computational-thinking-with-the-music-industry/>
- Rose, D. (2000). Universal design for learning. *Journal of Special Education Technology*, 15(3), 45–49.
- Shah, M. (2019). Scaffolding and Assessing Teachers' Examination of Games for Teaching and Learning. In D. Ifenthaler & Y. J. Kim (Eds.), *Game-Based Assessment Revisited* (pp. 185–210). Springer International Publishing. https://doi.org/10.1007/978-3-030-15569-8_10
- Shute, V. J. (2011). Stealth assessment in computer-based games to support learning. *Computer Games and Instruction*, 55(2), 503–524.
- Shute, V. J., Smith, G., Kuba, R., Dai, C.-P., Rahimi, S., Liu, Z., & Almond, R. (2021). The Design, Development, and Testing of Learning Supports for the Physics Playground Game. *International Journal of Artificial Intelligence in Education*, 31(3), 357–379. <https://doi.org/10.1007/s40593-020-00196-1>
- Shute, V. J., & Sun, C. (2019). Games for Assessment. In J. L. Plass, R. E. Mayer, & B. D. Homer (Eds.), *Handbook of Game-Based Learning* (pp. 491–512). MIT Press.
- Shute, V. J., Wang, L., Greiff, S., Zhao, W., & Moore, G. (2016). Measuring

- problem solving skills via stealth assessment in an engaging video game. *Computers in Human Behavior*, 63, 106–117.
- Steinkuehler, C., & Squire, K. (2014). Video Games and Learning. In *The Cambridge Handbook of the Learning Sciences* (2nd ed., pp. 377–396). Cambridge University Press.
- Thanapornsangst, S., Zheng, Y., & Holbert, N. (2020). Understanding Youth's Personal Connection to Data in a Gamelike Assessment System Through Learning Analytics and Qualitative Analysis. *Proceedings of the 2020 Connected Learning Summit*.
- Tufte, E. (1983). *The visual display of quantitative information*. Graphics Press.
- Turkle, S., & Papert, S. (1990). Epistemological Pluralism: Styles and Voices within the Computer Culture. *Signs: Journal of Women in Culture and Society*, 16(1), 128–157. <https://doi.org/10.1086/494648>
- Vahey, P., Rafanan, K., Patton, C., Swan, K., van 't Hooft, M., Kratcoski, A., & Stanford, T. (2012). A cross-disciplinary approach to teaching data literacy and proportionality. *Educational Studies in Mathematics*, 81(2), 179–205. <https://doi.org/10.1007/s10649-012-9392-z>
- Wachter, S., Mittelstadt, B., & Floridi, L. (2017). Why a right to explanation of automated decision-making does not exist in the general data protection regulation. *International Data Privacy Law*, 7(2), 76–99.
- Walker, S., & Hutchison, L. (2021). Using Culturally Relevant Pedagogy to Influence Literacy Achievement for Middle School Black Male Students. *Journal of Adolescent & Adult Literacy*, 64(4), 421–429. <https://doi.org/10.1002/jaal.1114>
- Wertsch, J. V. (1985). *Vygotsky and the Social Formation of Mind*. Harvard University Press.

Wiliam, D. (2018). *Embedded formative assessment* (2nd ed.). Solution Tree Press.

Zheng, Y., Blikstein, P., & Holbert, N. (2020). Combining Learning Analytics and Qualitative Analysis for the Exploration of Open-Ended Learning Environments. *Proceedings of Constructionism 2020*.

ABOUT THE ETC PRESS

The ETC Press was founded in 2005 under the direction of Dr. Drew Davidson, the Director of Carnegie Mellon University's Entertainment Technology Center (ETC), as an open access, digital-first publishing house.

What does all that mean?

The ETC Press publishes three types of work: peer-reviewed work (research-based books, textbooks, academic journals, conference proceedings), general audience work (trade nonfiction, singles, Well Played singles), and research and white papers

The common tie for all of these is a focus on issues related to entertainment technologies as they are applied across a variety of fields.

Our authors come from a range of backgrounds. Some are traditional academics. Some are practitioners. And some work in between. What ties them all together is their ability to write about the impact of emerging technologies and its significance in society.

To distinguish our books, the ETC Press has five imprints:

- **ETC Press:** our traditional academic and peer-reviewed publications;

- **ETC Press: Single:** our short “why it matters” books that are roughly 8,000-25,000 words;
- **ETC Press: Signature:** our special projects, trade books, and other curated works that exemplify the best work being done;
- **ETC Press: Report:** our white papers and reports produced by practitioners or academic researchers working in conjunction with partners; and
- **ETC Press: Student:** our work with undergraduate and graduate students

In keeping with that mission, the ETC Press uses emerging technologies to design all of our books and Lulu, an on-demand publisher, to distribute our e-books and print books through all the major retail chains, such as Amazon, Barnes & Noble, Kobo, and Apple, and we work with The Game Crafter to produce tabletop games.

We don't carry an inventory ourselves. Instead, each print book is created when somebody buys a copy.

Since the ETC Press is an open-access publisher, every book, journal, and proceeding is available as a free download. We're most interested in the sharing and spreading of ideas. We also have an agreement with the Association for Computing Machinery (ACM) to list ETC Press publications in the ACM Digital Library.

Authors retain ownership of their intellectual property. We release all of our books, journals, and proceedings under one of two Creative Commons licenses:

- **Attribution-NoDerivativeWorks-NonCommercial:** This license allows for published works to remain intact, but versions can be created; or

- **Attribution-NonCommercial-ShareAlike:** This license allows for authors to retain editorial control of their creations while also encouraging readers to collaboratively rewrite content.

This is definitely an experiment in the notion of publishing, and we invite people to participate. We are exploring what it means to “publish” across multiple media and multiple versions. We believe this is the future of publication, bridging virtual and physical media with fluid versions of publications as well as enabling the creative blurring of what constitutes reading and writing.