FLORIDA STATE UNIVERSITY

ANNE SPENCER DAVES COLLEGE OF EDUCATION, HEALTH, AND HUMAN SCIENCES

GAME-BASED ASSESSMENT OF STATISTICAL SELF-EFFICACY: AN ALTERNATIVE TO THE SELF-REPORT OF INTERNAL UNOBSERVABLE BELIEFS.

By

G. CURT FULWIDER

A Dissertation submitted to the
Department of Educational Psychology and Learning Systems
in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

2026

G. Curt Fulwider defended this dissertation on June 16, 2026.
The members of the supervisory committee were:

Bret Staudt Willet
Professor Directing Dissertation

Colleen Ganley
University Representative

Jeannine Turner
Committee Member

Russell Almond
Committee Member

The Graduate School has verified and approved the above-named committee members, and certifies that the dissertation has been approved in accordance with university requirements.

ACKNOWLEDGMENTS

> Anyway, I keep picturing all these little kids playing some game in this big field of rye and all. Thousands of little kids, and nobody's around—nobody big, I mean—except me. And I'm standing on the edge of some crazy cliff. What I have to do, I have to catch everybody if they start to go over the cliff... That's all I'd do all day. I'd just be the catcher in the rye and all.
>
> (Salinger, 1951, p. 224)

A dissertation is the accumulated patience, sacrifice, feedback, encouragement, and teaching of an entire network of people quietly carrying the author to the finish line. Our work stands on the shoulders of giants, but our lives rest on those willing to help carry us forward. I am grateful to all of those who have supported me in this work, and I would like to take a moment to acknowledge some of the most important people in my life.

Wenting. My wife, Lumi's mom, and my best friend. I never thought I could love anyone more until we had Lumi. Everything good that has happened to me traces back, in one way or another, to the night we met while launching lanterns. Left to my own devices, I know I would never have come this far. You have sacrificed so much to join me in this insane endeavor, and forever I will be yours. What is frightening right now is that we are done and now have to decide what comes next. It is scary, but also deeply exciting. We are about to start a new chapter in our lives, like the start of a new month. So, rabbit, rabbit, rabbit.

Lamoreaux & Gerry. My parents gave me the foundation for everything that followed. Mom provided the creativity, passion, and reckless disregard for societal norms that taught me "to thine own self be true." Dad provided the analytics, wonderment, and intellectual curiosity that form the most basic foundation of an academic life. Together, you gave me a home filled with love and support. You trusted me as an adult from the start, while also giving me a childhood mercifully unaware of how hard life can really be.

宋冬来和刘俊兰。 作为雯婷的父母，感谢你们不仅接纳我成为家人，也在我们追求这个重要人生目标的过程中始终给予耐心与支持。你们愿意远渡重洋来到这里陪伴和帮助我们，这份付出令人感动。这不仅仅是责任，更是你们对女儿、对我们这个小家庭最深厚的爱。最重要的是，在露米年幼的那几年里，正因为有你们帮忙照顾她，我们才能够完成这项工作。

Jeannine (The Shadow Advisor). You were a committee member, but also my shadow advisor and a dear friend. You had the greatest skepticism about my work, which meant you also gave the most valuable feedback. Cognitively and non-cognitively, you pushed this dissertation to the quality it has now. Most importantly, it was in your class that I first learned about self-efficacy. That class, and the project that came out of it, became the catalyst for my dissertation topic.

Bret (The Advisor). You were the first realization that I could keep my earring and still be a successful academic. You found me at Black Dog stuck and floundering, with no real hope that I was going to survive graduate school. It turns out I was not that far from shore, and the water was not that deep, but when I was underwater there was no way to see the coast. Thank you for pulling my head above water and reminding me that I can swim just fine.

Val. Yes, stealth assessment, *Physics Playground*, and the rest of that line of work are central to my current research. But I suspect I would have found the work wherever I was. What you showed me was that an academic can have high standards and real drive without losing a deep love and kindness for the people around them. More specifically, you showed me how to be a diligent and insightful researcher and, most importantly, how to present myself logically, calmly, and clearly, even if I am still working on that last one.

Russell. Thank you for your patience with my questions and for your generosity with your knowledge. You are likely the only person who fully understood my drive to do research for the sake of research rather than for any extrinsic motivation. You are an unending font of knowledge, an anomaly in the garden where, no matter which fork in the path one chooses, you are there providing the water of life for anyone willing to stop and drink. Thank you for letting me stop by now and then.

Sangeeta. When I started at Censio, I assumed it would be a quick job before moving on to the next thing. That had been my working experience up to that point. What followed instead were years of intense design, development, and fun. You brought together an incredible team of people and created a culture that deserves to be studied for its success. On a personal level, your earnest support was a major factor in my own progress. Not only did I benefit from your advice and experience, but many of the techniques I learned while working at Censio became tools I could repurpose in this dissertation. Without them, I have little doubt that my results would have been less than stellar.

Alysia. You landed in positions of increasing responsibility and influence at the same moments I needed compassion and flexibility. My journey through graduate school has been less than tra-

ditional or linear, and you provided the open-mindedness and understanding that allowed me to continue my work in face of challenges. I am grateful for your support and friendship.

Bryan. Your patience and concern are the main reasons I survived the forms, the decisions, and the business side of completing a doctorate. Through my absentmindedness and general cluelessness, we somehow also became great friends. We watched our floor fall silent during COVID and then slowly come back to life. Throughout all of that, you held the floor together, and yours was the one office that never changed and was never without a friendly, if occasionally scolding, smile and hug.

The Educational Psychology and Learning Systems Department. For the past ten years, EPLS has been my home. The hallways of Stone Building's third floor are as familiar to me as my own. A great deal has changed, however, and it feels more foreign each day I am away. I was fortunate to begin before the last great legacies of the field retired, and I am fortunate to have been present for the next great legacy. I stumbled into this department by pure happenstance: joining ISLT for the technology, gaining a deeper understanding of educational psychology through Learning and Cognition, and then finding my truest passion in Statistics and Measurement. I leave feeling accomplished, intellectually nourished, and finally ready to do something worthwhile with everything I have learned. Even as it changes, EPLS will always be my home.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

Self-efficacy is a consequential self-belief in life and especially in educational settings, yet it is often measured through self-report despite being dynamic, context-sensitive, and vulnerable to response bias. My dissertation investigated whether behavioral proxies derived from gameplay could predict students' statistical self-efficacy beliefs without relying exclusively on self-report. To address this problem, I designed *Mean Alchemy*, a learning game built to elicit behaviors theoretically related to self-efficacy, including persistence after failure, risk-taking, and goal setting. The study used a nonexperimental predictive design grounded in Evidence-Centered Design (ECD) and educational data mining. Pretest and gameplay data were available for 86 students for direct model comparisons. A total of 480 model configurations were evaluated across four model families—Logistic Regression, Logistic Regression with L1 regularization, Naive Bayes, and Gradient Boosting—to predict grouped pretest and posttest self-efficacy outcomes.

Results showed a modest but consistent predictive signal for pretest self-efficacy beliefs across multiple reasonable model families. Agreement among models was more informative than the performance of any single algorithm in isolation. The strongest recurring predictors included quest difficulty, attempts after failure, and interest in math and statistics. Posttest self-efficacy was less predictable than pretest self-efficacy, suggesting that later belief states were more dynamic and less recoverable from participant-level aggregated gameplay summaries. Confusion-matrix follow-up analyses further indicated that the hardest cases to classify were often those showing changes in self-efficacy between pretest and posttest.

These findings support the feasibility of using stealth assessment to infer self-efficacy from observable behavior in a game-based environment. More broadly, the study contributes a design process for engineering proxy features alongside gameplay rather than mining them retrospectively from existing environments. The results also suggest that future work should move toward finer-grained temporal modeling to better capture changes in self-efficacy as they unfold during performance.

CHAPTER 1

INTRODUCTION

1.1 Why Self-Efficacy Matters

Regardless of a person's true capabilities and knowledge, if they lack belief that their skills and efforts will suffice to obtain their goal, they are unlikely to persist or tackle greater challenges (Bandura, 1997). Self-efficacy is pivotal in understanding motivation and achievement (Schunk & Pajares, 2002). Defined as the belief one holds about their ability to manage and execute actions required to achieve specific outcomes, self-efficacy plays a critical role in how people think, behave, and feel (Bandura, 1986, 1995, 1997). Constructs like self-efficacy, self-concept, and anxiety are internal self-beliefs as opposed to objective knowledge or skills that are the focus of typical educational programs (e.g., math, science, reading). Self-efficacy is one of the strongest self-concept constructs in predicting academic achievement (Stankov & Lee, 2014). This dissertation responds to that problem by developing and evaluating a stealth-assessment approach to self-efficacy embedded within a learning game so that students' self-efficacy can be inferred primarily from gameplay behavior and educational assessments can be interpreted more fairly and accurately.

Self-efficacy is more difficult to measure than objective constructs because it is a fluid state that can fluctuate from moment to moment, solidifying as the individual grows with positive experience and mastery or backsliding with negative experience (Bandura, 1977). The only established method for assessing self-efficacy is through self-report (e.g., surveys, interviews). Self-report makes assessments of self-efficacy susceptible to bias (Kormos & Gifford, 2014; Watson et al., 2006) and limits how frequently efficacy beliefs may be measured due to testing fatigue and respondent burden (Ben-Nun, 2008). This is especially problematic for researchers and practitioners who want to understand how self-efficacy changes over time, in response to interventions, or in different contexts. For example, a student may feel confident in their math abilities at the beginning of the school year but may experience a decline in self-efficacy after receiving a poor grade on a test. Without frequent and accurate measurements of self-efficacy, it is difficult to identify when and why these changes occur and how to intervene effectively.

Furthermore, the confounding influence of efficacy beliefs on knowledge- and skill-based assessments magnifies the importance of including assessments of self-efficacy alongside assessments of knowledge or skills. For instance, tests and quizzes are the most common form of assessment in U.S. middle and high schools and carry substantial weight in students' grades (Banilower et al., 2013; Guskey & Link, 2019). Testing anxiety is a common issue that can influence a student's ability to perform on a test in a way that reflects their actual knowledge (von der Embse et al., 2018). Testing anxiety is often a symptom of weak self-efficacy (Bandura, 1995; Roick & Ringeisen, 2017). Thus, the influence of internal self-beliefs (e.g., self-efficacy) undermines the accuracy of knowledge-based assessments by interfering with a student's ability to perform on an assessment—a bias from self-belief. That means assessments that do not consider self-efficacy cannot differentiate between students in need of remediation (i.e., lacking competence) and those who possess competencies "overruled by self-doubt" (Bandura, 1997, p. 37).

As our understanding of human learning and our available technologies improve, there is an ongoing discussion about how assessments should be used and what alternative methods should be developed (Black & Wiliam, 2010; Delandshere, 2002; Mislevy, 1993; Pellegrino et al., 1999; Shute & Becker, 2010). Reducing bias is an important goal of assessment improvement. Whether the focus is on racial, gender, or other subgroup biases (Hutchinson & Mitchell, 2019), assessment methods (Kates et al., 2022), or prior traumatic experiences—for example, negative experiences in life or schooling that may influence a person's ability to complete an assessment (Fraine & McDade, 2009)—addressing bias means tackling unaccounted-for variables that have an unintended and unfair influence on students' scores. Most biases, like the examples given, are external. Internal biases, those that come from within the individual, have too often been overlooked. While typical methods of assessment (e.g., self-report, situational-judgment tasks) can measure a student's efficacy beliefs, they are not without bias, and none are effective for continuous administration—which is needed to capture a complete picture of the dynamic nature of self-efficacy (Bandura, 1997).

1.2 Purpose and Research Questions

With this context in mind, the purpose of my study was to develop and evaluate a stealth assessment of self-efficacy embedded within a learning game, aiming to provide a largely unobtrusive, theory-informed estimate of students' self-efficacy and thereby enhance the fairness and accuracy of educational assessments. To accomplish this purpose, I proposed an assessment of self-efficacy that

relied on data mined from a player's gameplay experience, supported inferences based on authentic behaviors, and reduced reliance on self-report.

More specifically, this dissertation uses a quantitative, nonexperimental predictive design. It is grounded in educational data mining and tests whether theory-informed behavioral proxies drawn from gameplay could predict students' self-efficacy beliefs. The study focuses on three broad families of observable behaviors—persistence, risk-taking, and goal setting—along with attitudinal proxy variables that help distinguish self-efficacy from adjacent influences.

This dissertation addresses one research question: **To what extent can behavioral proxies of self-efficacy in a game-based assessment predict self-efficacy beliefs?**

To answer that question, the study used gameplay telemetry and survey data from a digital learning game to evaluate whether self-efficacy can be inferred from observable behavior rather than self-report alone. The broader outcome was to determine whether a stealth assessment of self-efficacy can produce interpretable, theory-aligned evidence that supports fairer assessment practice and more accurate interpretation of student performance.

CHAPTER 2

THEORETICAL FRAMEWORK

I approach this work from the theoretical viewpoint of Albert Bandura's self-efficacy (Bandura, 1997). Skinner (1976), most closely associated with radical behaviorism, argued that psychological research should focus exclusively on observable behavior, treating the unobservable inner workings of the mind as unnecessary for scientific analysis—often characterized as a "black box". Instead, he argued that behavior is shaped by environmental contingencies and can be systematically measured and modified through reinforcement. This view positioned individuals as reactive entities, whose actions were shaped by external factors rather than internal self-beliefs and related mental processes.

Bandura (1995), while not entirely opposed to Skinner's theories, offered a different perspective by proposing an "agentic" perspective that emphasized the role of internal processes and beliefs—the unobservables within the black box—in human action. Bandura argued that these internal determinants, although dismissed by radical behaviorism, play a crucial role in shaping behavior and should not be overlooked. He introduced the concept of self-efficacy and emphasized its central role in understanding human motivation and behavior.

Bandura's agentic perspective suggested that human motivation is not merely a response to external stimuli but the result of a triadic relationship between the person, their behavior, and the environment (Bandura, 1986). Within this framework, self-efficacy emerges as a key self-belief that influences how individuals approach and persist in tasks. While much of Bandura's early work focused on overcoming phobias, later work has highlighted the significance of self-efficacy in academic achievement (Bandura, 1995; Stankov & Lee, 2014).

Despite these advances in understanding the role of self-beliefs, modern educational testing remains largely focused on the measurement of competency. Neglecting learners' self-belief factors may introduce biases in competency assessments, particularly in cases where students' self-efficacy beliefs significantly influence their performance (i.e., behaviors), regardless of their actual competence (Bandura, 1997). As such, there is a need to update assessment practices to incorporate our understanding of self-efficacy and other self-beliefs that influence learning outcomes.

When given a task—one whose completion is expected to produce a desired outcome—a person makes a subjective and unconscious evaluation of their existing capabilities, knowledge, skills, and

related resources that forms a belief about the likelihood of completing the task. The evaluation can be influenced by physiological responses (e.g., nausea before a test), verbal persuasion from oneself or another person, observing peers accomplish the same task, and—strongest of all—the experience of mastering or failing the task for oneself (Bandura, 1997). There is no "filter" distinguishing accurate from inaccurate information in the formation of efficacy beliefs. However, a range of related constructs (e.g., self-concept, self-esteem) also influence one's choice of action—that is, motivation (Schunk & Pajares, 2002). Thus, regardless of talent, skill, acclaim, or achievement, individuals with extensive mastery experience may still struggle to perform—and fail to effectively organize their capabilities—if their self-beliefs are undermined by context or feedback (Bandura, 1997).

Self-efficacy is a core driver of motivation and achievement (Bandura, 1997; Zimmerman, 2000). Although there is a complex network of internal constructs (e.g., self-concept, goals), self-efficacy seems to be a key construct in predicting achievement (Stankov & Lee, 2014; Stankov et al., 2014). However, if efficacy beliefs are not a pure reflection of objective capabilities, then individuals with weak self-efficacy beliefs for a task—even if they are fully capable of completing the task—are likely to underperform by giving up early, being distracted by physiological symptoms, or even refusing to start. Thus, performance on an exam or presentation used to assess learning by having students demonstrate their knowledge is largely impacted by efficacy beliefs that do not necessarily align with the individual's true competence.

CHAPTER 3

LITERATURE REVIEW

This literature review supports a quantitative, nonexperimental predictive study in which gameplay telemetry is used to estimate self-efficacy beliefs without interrupting the learning experience. The review therefore builds toward three connected outcomes: a theory-grounded rationale for using persistence, risk-taking, and goal setting as behavioral proxies (i.e., ECD); a design rationale for embedding those opportunities within a game (i.e., Stealth Assessment); and an analytic rationale for using educational data mining to evaluate whether those proxies produce a stable predictive signal across models. To make that logic explicit, the review first establishes the assessment-design foundation in ECD and stealth assessment, then explains the limitations of relying on ECD alone, then turns to educational data mining as a complementary analytic approach, and finally narrows to prior work on game-based assessment, measurement bias, and the observable correlates of self-efficacy used in this study.

3.1 Evidence-Centered Design

ECD is a systematic framework for developing assessments that link observable evidence to claims about learner competence through structured evidentiary reasoning (Mislevy et al., 2003). ECD emphasizes creating structured opportunities for learners to demonstrate observable behaviors that are evidence of their mastery of knowledge, skills, and other constructs (DiCerbo, 2014; Lee & Recker, 2017; Shute et al., 2021).

For example, DiCerbo (2014) used an ECD approach to create a game-based assessment of task persistence within a digital learning game. Within the game, students completed quests to progress to other areas and earn rewards. The game was an existing commercial game and the researchers designed the assessment by determining what in-game behaviors would serve as evidence of that construct based on established literature on task persistence. The fine-grained telemetry data collected from students' gameplay (i.e., the logs of their in-game actions) provided the observable evidence that the researchers then used to make inferences about students' task persistence.

Lee and Recker (2017) used log data from learning management systems to measure students' self-regulated learning strategies. By applying ECD principles, they identified specific online behaviors (e.g., frequency of accessing course materials, time spent on tasks) that could serve as evidence of self-regulated learning. This approach allowed them to make inferences about students' learning strategies based on their interactions with the digital learning environment.

By grounding assessment design in theory and explicitly linking observable actions to claims about underlying competencies and self-beliefs, this study extends ECD, which has traditionally emphasized competence. This extension ensures alignment between tasks and both what learners can do and the beliefs that shape whether those capabilities are enacted. This theory-driven approach supports the development of valid and reliable assessments by connecting task design to established principles in educational psychology and measurement theory (Mislevy et al., 2003).

In this study, ECD guides the process of identifying key behavioral indicators of self-efficacy, such as persistence, risk-taking, and goal-setting. It also structures in-game opportunities through which players can demonstrate these behaviors (Bandura, 1997; Bouffard et al., 2005; Krueger & Dickson, 1994; Rosenthal et al., 1991; Zimmerman, 2000). Game challenges are designed to elicit observable evidence of self-efficacy—such as persisting on difficult tasks or engaging in skill-based risks—intentionally embedded within the game environment. These behaviors are grounded in self-efficacy theory, which posits that individuals' beliefs about their capabilities influence their motivation and actions (Bandura, 1997). Through this application of ECD, the game environment functions as a structured context in which player behaviors are systematically linked to internal constructs, generating meaningful data for assessment while preserving the integrity of the learning experience.

Using ECD as the basis to create an assessment of self-efficacy may resolve the limitations that prevent teachers, faculty, and other stakeholders from considering and supporting students beyond simple remediation of prerequisite knowledge. ECD provides the framework needed to ensure the behaviors being observed are theoretically connected to the target construct (i.e., higher persistence is associated with higher self-efficacy). Taking a behavioral approach then reduces dependence on self-report. Embedding the assessment deeply within a learning game makes the assessment unobtrusive and provides an environment for authentic behaviors to emerge within natural learning contexts. Improving authenticity and decreasing intrusiveness has the added benefit of minimizing the influence of testing anxiety and other confounding variables.

However, self-efficacy exists within a complex system of interrelated motivational constructs. The theory-driven approach relies on prior research aimed at creating generalizable conclusions. In other words, seminal work is limited by an ability to fully capture the complexity of human cognition and its influence on behavior. The behaviors and context associated with self-efficacy are intricate and multifaceted, making them unsuitable for reduction to simple, single observations. As a result, more advanced measurement methods, such as machine learning, are necessary to uncover and model these dynamic relationships. Furthermore, the environment in which these behaviors are observed must be rich and engaging enough to generate the detailed data required for analysis. This is where games become instrumental; they create immersive environments that naturally elicit meaningful behaviors. By embedding ECD within the design of games, stealth assessment becomes possible—allowing the collection of authentic data that aligns with theoretical constructs while remaining unobtrusive to the player's experience.

3.1.1 Stealth Assessment

Stealth assessment applies an ECD approach to embed assessment seamlessly within well-designed learning games (Shute, 2011) and provides the ideal framework for drawing theoretically sound connections between behavior and belief (Choi et al., 2023; Shute et al., 2021; Ventura & Shute, 2013). At the heart of ECD is the alignment of knowledge and skills—in this case, self-efficacy beliefs—with game tasks that will elicit observable evidence of players' efficacy beliefs. This framework allows for creating the game, tasks, and assessments in solid alignment with minimal (or no) interruption of learning in gameplay, allowing for a real-time, stealth assessment of students' learning progress and efficacy beliefs.

Stealth assessment, rooted in ECD, offers a method to unobtrusively measure self-efficacy through the analysis of in-game behaviors, also known as trace data and learning analytics. ECD ensures that the tasks within the game are designed to elicit observable behaviors that accurately reflect the players' efficacy beliefs, making it a suitable framework for this study. This approach is supported by the findings of Shute (2008) and Mislevy et al. (2003), who highlighted the benefits of formative assessment and behavioral-based assessments. By minimizing interruptions in gameplay, ECD facilitates real-time assessment, allowing for continuous monitoring and evaluation of self-efficacy without disrupting the learning process. Real-time, unobtrusive assessment also reduces reliance on repeated self-report measures, providing a more accurate and dynamic representation of a student's self-efficacy.

Prior studies have had success using ECD and stealth assessment to measure hard-to-observe constructs like computational thinking (Rowe et al., 2021), systems thinking (Shute et al., 2010), and creativity (Shute & Rahimi, 2021). These earlier works depend on an evidence model that uses theory to build connections between behaviors and the construct being measured. For example, making incremental changes and testing the result is evidence of problem solving (Liu & Israel, 2022). In the case of self-efficacy, the behaviors are outcomes of high or low self-efficacy. Like diagnosing an illness by symptoms, individuals who do not believe in their personal efficacy persist less, set lower goals for themselves, and avoid challenges (Bandura, 1997). These tendencies are the observable correlates to self-efficacy.

3.1.2 Limitations of Evidence-Centered Design

ECD can require a substantial investment of time and design effort, which can make it difficult to implement at scale. Designers must predefine the behaviors to be treated as evidence, and developing tasks that align closely with theoretical constructs often requires considerable manual work. Scaling ECD-based assessments across diverse populations and learning contexts can therefore become cumbersome, especially when the assessment depends on carefully engineered opportunities for evidence collection (Almond et al., 2020; DiCerbo, 2014; Lee & Recker, 2017; Mislevy et al., 2003).

ECD's reliance on predefined tasks may also oversimplify complex human behaviors. The approach can struggle to capture the full range of behaviors shaped by dynamic and multifaceted learning environments. Confounding variables may prevent students from displaying the expected behaviors, which can produce incomplete or less informative assessments. Human behavior, especially in learning contexts, is influenced by factors that cannot always be fully anticipated in a prescriptive design. Consequently, the richness and variability of student responses may be harder to capture within a tightly specified evidence model (Choi et al., 2023; Mislevy et al., 2003).

Designing for specific, observable behaviors also risks overlooking more nuanced actions that could be valuable indicators of learning. Subtle and emergent behaviors, which may be important for understanding constructs like self-efficacy, can go unnoticed when the evidence model is specified too narrowly in advance. In that sense, the complexity of human behavior often requires a more flexible analytic approach that can detect meaningful patterns after interaction data have been collected rather than only those anticipated at the design stage (Baker, 2019; Baker & Inventado, 2014; Long & Siemens, 2011).

3.2 Educational Data Mining

Educational Data Mining (EDM) offers a complementary response to the limitations of ECD by focusing on the extraction of meaningful patterns from educational data to better understand and improve learning processes (Baker & Inventado, 2014). The practice takes advantage of advanced algorithms and statistical techniques to analyze complex—often large-scale—sets of data generated by students' interactions with educational platforms (e.g., learning management systems, learning games). EDM provides an advantage over traditional assessment methods (e.g., quizzes) by offering insights into dynamic and complex progressions of behavior, allowing these behaviors to emerge naturally through the analysis of fine-grained data, rather than being constrained by predefined connections between competency and behavior. In educational games, EDM is especially valuable as it can process rich gameplay data, identifying meaningful sequences of behaviors that might indicate constructs like self-efficacy or engagement. The use of EDM in educational research has increased substantially, leading to more personalized and data-driven learning interventions (Baker, 2019).

In the context of learning games, EDM serves as a powerful tool for analyzing player behaviors and drawing inferences about self-belief constructs like self-efficacy (Ventura & Shute, 2013). The rich fine-grained data made available by digital learning games provide a real-time record of specific actions taken by a player, which EDM methods (e.g., machine learning algorithms) can organize into analyzable patterns. By processing these data streams, EDM can detect variations in player behavior, such as changes in persistence or risk-taking, that are linked to self-efficacy beliefs. This data-driven approach is particularly effective in identifying subtle behavioral cues that might not be explicitly accounted for in theory-driven designs. The flexibility of EDM allows researchers to adapt and refine measurement models as new patterns emerge, contributing to a more comprehensive understanding of how students interact with educational games. Ventura and Shute (2013) emphasize the value of EDM for unobtrusive assessments that maintain the immersive experience of the game, ensuring that learning remains engaging.

EDM offers several advantages over traditional assessment methods, particularly when measuring constructs in complex environments like educational games (Long & Siemens, 2011). Traditional assessments, like surveys or observation, often disrupt the learning experience and may fail to capture the real-time, dynamic nature of student behaviors. In contrast, EDM continuously monitors player interactions, providing a richer and more nuanced dataset for understanding student learning

processes. The scalability of EDM methods also means they can be applied to large datasets, making them ideal for analyzing educational games played by diverse populations. Another significant advantage is the ability of EDM to reveal emergent behaviors, which can lead to new insights that theory-driven methods might overlook. However, it is essential to interpret EDM results cautiously, as these data-driven discoveries still require validation against established theories to ensure their educational relevance (Long & Siemens, 2011).

3.2.1 The Benefits of EDM

EDM provides a scalable and adaptable way to analyze large datasets without the need for extensive manual design. It uses algorithms to extract patterns from complex data, making it more efficient for large-scale applications. EDM can adapt to diverse learning environments, offering insights that are not tied to a single predefined task. The approach allows researchers to explore new behaviors and relationships as they emerge, rather than being constrained by initial design assumptions. This scalability makes EDM well suited to analyzing rich, real-time data from educational games and adaptive learning technologies (Baker, 2019; Baker & Inventado, 2014; Long & Siemens, 2011).

EDM can uncover subtle and emergent behaviors that are often missed by traditional, theory-driven designs. By analyzing rich, fine-grained data from student interactions, EDM can reveal patterns that are not immediately apparent. This allows for the detection of nuanced behaviors, such as variations in persistence or engagement, which are valuable for assessing learning. EDM's flexibility is especially useful when the target construct is expressed through complex patterns of interaction rather than a single predefined behavior (Baker, 2019; Baker & Inventado, 2014; Ventura & Shute, 2013).

EDM complements theory-driven approaches by using empirical data to inform and refine assessments. Data-driven analysis can validate or challenge theoretical assumptions, offering a balance between theory and empirical evidence. EDM provides a more nuanced view of student behaviors and better accommodates the unpredictability of real-world learning contexts. This combination of data and theory allows assessment models to be iteratively improved while remaining connected to substantive interpretation (Choi et al., 2023; Long & Siemens, 2011; Mislevy et al., 2003).

3.2.2 Proposed Solution: Merging ECD and EDM

ECD will guide the design of the game, ensuring a strong theoretical foundation for observing key variables. The game will be structured around constructs like persistence, risk-taking, and goal setting. ECD will determine how opportunities are presented to students, creating meaningful scenarios for behavior observation. This theoretical framework helps align the game with established constructs, making the assessment more defensible. Using ECD for design sets the stage for meaningful data collection without compromising the educational goals of the game (Choi et al., 2023; Mislevy et al., 2003; Shute et al., 2021).

EDM will be the primary tool for analyzing gameplay data, offering a flexible and data-driven approach to measurement. It will identify patterns in player behavior, including subtle and emergent actions that theory alone might miss. The approach reduces the burden on designers to engineer every behavior by allowing data to reveal unexpected insights. EDM's analysis of rich gameplay data provides a fuller picture of how students engage with the game, helping the assessment capture the complexity of learning behaviors more effectively (Baker & Inventado, 2014; Long & Siemens, 2011; Ventura & Shute, 2013).

Combining ECD and EDM creates a balanced approach that leverages the strengths of both theory and data. ECD provides a structured framework for game design, while EDM offers flexibility in measurement and analysis. This hybrid model allows assessments to be both theory-informed and data-driven, improving interpretability while also supporting scalability. By using both approaches, the assessment can uncover subtle patterns while remaining grounded in established research. The result is a more comprehensive approach to measuring complex constructs like self-efficacy in educational environments (Baker & Inventado, 2014; Choi et al., 2023; Mislevy et al., 2003; Shute et al., 2021).

3.3 Game-based Assessments of Self-Beliefs

Choi et al. (2023) reviewed studies attempting to measure self-regulated learning using logfile data and found a consistent lack of theoretical justification for chosen behaviors meant to represent self-regulated learning. The authors identified a need for more rigorous "consideration of theories, contexts, and research questions..." in drawing the connection between behaviors and constructs (Choi et al., 2023, p. 81).

McQuiggan et al. (2008), for example, achieved up to 77% and 80% accuracy (AUROC) testing Naive Bayes and decision trees on two types of models—both firmly grounded in theory—for measuring self-efficacy. The first "static" model was trained on self-reports of efficacy beliefs, observable behaviors from an online tutoring system, and demographic data. The second "dynamic" model added biofeedback data (i.e., heart rate and galvanic skin response) to the static model. While this research was important, their approach had a number of limitations. For example, physiological data is intriguing to include, but intrusive and hard to scale. Including demographics (e.g., race, gender) in training data runs the risk of perpetuating systemic inequalities present in the surrounding cultural milieu (Cheuk, 2021; Kostick-Quenet et al., 2022). And finally, the models used in McQuiggan et al. (2008) were trained on data collected from self-report measures, thus their approach could perpetuate the proposed self-belief bias. More broadly, recent work has shown that some negative self-report well-being indicators may not demonstrate measurement invariance across intelligence levels, which raises an additional concern that self-report scores are not always directly comparable across groups (Czerwiński et al., 2025).

McQuiggan et al. (2008) used the "four major processes" that Bandura (1997) described as the channels through which self-efficacy beliefs influence human behavior to justify their approach. In this approach, the "symptoms" of weak efficacy beliefs can be used to infer efficacy beliefs by their presence or absence (e.g., a highly persistent student would be expected to have high efficacy beliefs). They also included a thorough reporting of their variable definitions (McQuiggan et al., 2008, p. 113). However, they did not include a clear line of theory connecting the chosen behaviors and self-efficacy. Certain behaviors they selected (e.g., "locational features") seem arbitrary, possibly because of a false dilemma in their approach. McQuiggan et al. (2008, p. 87) claim there are "two fundamental approaches to modeling self-efficacy". The first approach is to rely on expert opinion and literature review. The second approach is entirely data-driven (i.e., supervised machine learning). There is another approach that could merge the theory-driven and data-driven approaches that addresses the limitations in McQuiggan et al. (2008) and make an assessment of self-efficacy, free of self-report, feasible.

Although Bandura's recommended method of measuring self-efficacy is self-report (Bandura, 1997), the emergence of self-efficacy research came before the availability of educational data mining methods that can make sense of the complex dynamic relationship between self-belief and competency constructs. Coupled with the rich fine-grained data available from digital learning games, it

is possible that the "symptoms" of weak efficacy beliefs (e.g., giving up early, avoiding challenges, setting low goals) could be behavioral proxies for estimating efficacy beliefs.

3.4 Measurement Challenges and Bias in Assessment

Currently, the most common method for assessing self-efficacy beliefs is via survey. Each question asks the respondent to rate, on a 100-point scale, their confidence in successfully completing a variety of tasks with varying difficulties—without necessarily completing the tasks (Bandura, 2006, 2012; DiBenedetto & Schunk, 2022). Situational judgment tests (SJTs) are also an option for assessment. In SJTs, a student is given relevant scenarios and asked—usually via multiple choice—to state how they would most likely react (Kyllonen, 2020). Because self-efficacy resides within the unobservable interior of the "black box", there is a common assumption that the only way to measure self-efficacy is via self-report (e.g., interviews, surveys, SJTs).

Self-report is notorious for having both self-serving and desirability biases (Kormos & Gifford, 2014; Watson et al., 2006). For instance, Kormos and Gifford (2014) found that when asked about potentially sensitive internal beliefs, participants were more inclined to misreport for fear of revealing weakness or to try and meet perceived expectations of the instructor. Others argued that it is impossible to completely remove the biases inherent in self-report (Lira et al., 2022; Rosenman et al., 2011). Therefore, addressing a self-belief bias cannot be dependent on a solution that introduces more bias than it resolves. Assessments based on observable authentic behaviors from the participant provide a better—potentially more accurate—measurement of competency and self-belief constructs (Mislevy et al., 2003).

Research has shown that it is possible to overcome biases in self-report by focusing on behaviors indicative of strong or weak self-beliefs. For example, in their study, Zhou and Winne (2012) compared self-reports of goal orientation with observed behaviors. Participants were given passages of text and asked to label passages from a given library of labels. The behavior being observed was the selection of a phrase that represented their goal orientation when studying a digital text. The observed behaviors (e.g., which labels were chosen) were better able to predict achievement than the students' self-reported goal orientations.

Self-efficacy, on the other hand, is more complicated. There is an ongoing series of decisions and influences occurring within students' minds that can impact behavior, one of which is knowledge or skill. The ongoing evaluation of the task considers the environment, the audience, the purpose

of the task, and related factors. Using the rich data collected from gameplay, a data-driven (i.e., exploratory) approach is not only feasible but could also be more accurate in identifying behaviors—or series of behaviors—that are not easily deduced from theory alone. Put simply, the complex data collected from gameplay is too rich for a single human—or even a panel of experts—to fully comprehend and requires the power of machine learning to identify.

It is possible that external behaviors (e.g., persistence) can be used as proxy behaviors for inferring efficacy beliefs. Through years of investigation into how efficacy beliefs influence observable behaviors, Bandura (1997) found that the key determinant of whether an individual—regardless of acquired skills—initiates the actions needed to accomplish a goal is the person's belief in the potential for success (i.e., self-efficacy). However, reliance on a single behavior (e.g., persisting a few minutes longer than average) is insufficient evidence (Almond et al., 2020). The body of behavioral evidence needed to make a valid inference of efficacy beliefs is also likely too complex to be interpreted by a human researcher or instructor alone (Denovan et al., 2023). A strong theoretical foundation is therefore required for building an assessment that can uncover complex patterns of behavior indicative of efficacy beliefs (Choi et al., 2023).

3.5 Observable Correlates to Self-Efficacy

Self-efficacy has a well-researched correlational relationship with many observable constructs like persistence on a difficult task, the setting and monitoring of goals, and skill-based risk-taking. Persistence has a directly observable proxy built into the definition (e.g., time-on-task or number of attempts; (Eisenberger, 1992; Feather, 1962)) and has repeatedly been shown to be strongly correlated with one's efficacy beliefs—that is, higher self-efficacy results in increased persistence in the face of difficulty and failure (Bandura, 1997; Lent et al., 1984; Liao et al., 2014; Ventura & Shute, 2013; Wu et al., 2020).

A strong sense of personal efficacy is also related to increased risk-taking (Beghetto, 2009; Deitzer et al., 2021; Krueger & Dickson, 1994; Rosenthal et al., 1991), which is observable as a willingness to risk reward or failure on relevant skill-based tasks (e.g., accepting a bonus challenge that risks losing in-game points). Goal setting and monitoring can be revealed by allowing participants the opportunity to record their goals and the steps they take to monitor their progress. Individuals who possess a stronger sense of personal efficacy tend to set higher goals and monitor their progress more frequently compared to their less self-efficacious peers (Bandura & Locke, 2003;

Gonzalez-DeHass et al., 2022; Huang, 2016; Krueger & Dickson, 1994; Wood & Bandura, 1989). Therefore, there are potential observable constructs that may help reveal students' efficacy beliefs.

3.5.1 Persistence

Feather (1962) provided a thorough exploration of persistence, the definition of which and means of measurement have remained virtually unchanged since:

> The general paradigm of the persistence situation is that in which a person is confronted with a very difficult or insoluble task and is unrestricted in either the time or number of attempts he can work at it. He is unsuccessful at each of these attempts at the task, but can turn to an alternative activity whenever he wishes. Persistence may be measured by the total time or total trials which the person works at the task before he turns to the alternative activity (Feather, 1962, p. 94).

While Feather (1962) reviewed work that treated persistence as either a trait or a construct, Eisenberger (1992) reported strong evidence to support the learnability of persistence—which Eisenberger called "industriousness" rather than persistence, but the two terms are similar in definition. By reinforcing greater effort on high-effort tasks, Eisenberger found that "the individual becomes more likely to exhibit high effort in the future" (Eisenberger, 1992, p. 248). Based largely on these two studies, I define persistence as continued effort—in terms of either number of attempts or time on task—towards a specified goal in spite of reasonable difficulty.

Wu et al. (2020) explored the influence of subjective task value on effort and persistence using survey data from 163 undergraduate students enrolled in engineering classes. Their results showed self-efficacy to be significantly correlated with persistence, $r(163) = .35$, $p < .01$. Further analysis using hierarchical regression revealed that self-efficacy significantly predicted persistence in engineering tasks, $\beta = 0.32$, $SE = 0.08$, $p < .001$. The addition of self-efficacy to the model explained an additional 10% of the variance in persistence, $\Delta R^2 = .10$, $\Delta F(1, 159) = 18.57$, $p < .001$. However, when the task value variables were added in Step 3, none of the variables, including self-efficacy, remained significant predictors of persistence, suggesting that the perceived value of the task is crucial for persistence.

Ventura and Shute (2013) demonstrated the validity of a stealth assessment of persistence. While self-efficacy is an internal belief, persistence is an observable behavior. Therefore, assessing persistence as a behavioral correlate of self-efficacy is essential for accurately evaluating self-efficacy. Ventura and Shute (2013) conducted a study involving 154 students who played a challenging

physics video game called Newton's Playground for approximately four hours. Their game-based assessment of persistence (GAP) used time spent on solved and unsolved physics problems. Using an external Performance Measure of Persistence (PMP) similar to the one Eisenberger (1992) used, Ventura and Shute (2013) validated their GAP. The GAP was positively correlated with the PMP ($r = .51$, $p < .01$ for low performers; $r = .22$, $p < .05$ for high performers), indicating that the game-based measure of persistence was consistent with the external measure.

Ventura and Shute (2013) also observed differences between high and low performers in their study. In Newton's Playground, players earned gold trophies by solving levels with minimal use of objects, while silver trophies were earned with moderate use of objects. They found that both high and low performers showed significant correlations between GAP and PMP measures, but the strength of these correlations differed. Low performers, defined as those in the bottom 50th percentile on gold trophies, had stronger correlations between GAP and PMP measures, suggesting that persistence is more evident and measurable among students facing greater challenges. High performers, defined as those in the top 50th percentile, exhibited significant but weaker correlations between GAP and PMP, indicating that these students did not need to exert as much persistence because they experienced a lower level of challenge. This reinforces the idea that a measure of persistence requires tasks to be sufficiently challenging to reflect the effort and perseverance of individuals accurately. Without sufficient challenge, the assessment may not effectively capture variations in persistence, which may be critical for evaluating self-efficacy.

3.5.2 Risk-taking

Risk-taking involves engaging in behaviors that have the potential for both significantly positive and negative outcomes. It is an essential aspect of decision-making and learning because it reflects individuals' willingness to face uncertainty in challenges. Higher levels of self-efficacy are generally associated with increased risk-taking because individuals with high self-efficacy focus on opportunities rather than threats and are confident in their ability to avoid negative outcomes (Beghetto, 2009; Deitzer et al., 2021; Krueger & Dickson, 1994; Montford & Goldsmith, 2016). Definitions of "risk-taking" are often context specific—much like self-efficacy is context specific—and focus on a willingness to engage in task-relevant behaviors with both potentially positive and negative outcomes that are uncontrollable or uncertain but dependent on skill.

Uncontrollability is inherent in risk-taking. However, there is a difference between tasks that are based entirely on chance (e.g., gambling) and those that depend on skill or ability. Congdon

et al. (2013) demonstrated that tasks with clearly different levels of controllability activate different regions of the brain. While being monitored using an fMRI, participants played one of two games. The first game involved randomly drawing from a pool where the distribution of objects was known (i.e., drawing marbles from a clear glass vase). The second game had a similar mechanic without the information needed to determine the chance of failure—much like a slot machine. While preliminary, their results suggest that less ambiguity was associated with increased activation in the valuation network during risky decisions, hinting that participants experiencing less ambiguity engaged in more deliberate effort to optimize their payout. Thus, game tasks used for the assessment of self-efficacy must rely on skill over chance to encourage thoughtful engagement rather than mindless clicking.

The results from Congdon et al. (2013) were demonstrated earlier by Krueger and Dickson (1994) without the use of fMRI. In a 2×2 experimental design, participants—153 business majors—completed business decision-making tasks that involved controllable dilemmas and uncontrollable "gambles" and received non-contingent positive or negative feedback. Path analysis results indicated that the influence of self-efficacy on decision-making behaviors was fully mediated by perceptions of opportunity and threat. While the influence on decision-making behavior was significant in both the dilemma and gambling contexts, the influence of opportunity/threat perceptions on behaviors in the dilemma context ($\beta = .341$ for opportunity, $\beta = -.453$ for threat) was stronger than in the gambling context ($\beta = .195$ for opportunity, $\beta = -.236$ for threat). Furthermore, the authors found that feedback on gambles did not influence behaviors on dilemmas and vice versa.

Similar to how a game designed to measure persistence must optimize the level of difficulty to create an opportunity for the player to persist, a game designed to measure risk-taking comes with constraints. First, a balance of opportunity and threat must be present. Without the opportunity for reward, there is no justification for risk, and vice versa. Second, the game task must limit the influence of chance; that is, the probabilities of either positive or negative outcomes must be close enough to depend on skill more than chance. These constraints are intended to create the ideal environment in which players can demonstrate behaviors indicative of their efficacy beliefs.

Assessing risk-taking as a measure of self-efficacy. Risk-taking, like self-efficacy, is domain specific (Congdon et al., 2013; Krueger & Dickson, 1994). Thus, evidence supporting the influence of self-efficacy on risk-taking behaviors should cover a variety of contexts. Beneath the decision to accept or reject a risky task is an ongoing judgment of opportunity and potential cost. When ambiguity is minimized, the judgment of risk and reward is based on the decision-maker's

perception of ability—that is, those with higher self-efficacy are more willing to undergo greater risk because their beliefs in personal efficacy reassure them that there is a greater chance of success.

Merritt and Tharp (2013), for example, conducted a quantitative cross-sectional study with 277 parkour and free-running practitioners to assess the effect of self-efficacy on the relationship between personality traits and risk-taking behaviors. The researchers used hierarchical multiple regression and mediation analysis. They found that greater reckless risk-taking behaviors appeared to be associated with higher levels of neuroticism ($p = .013$) and lower conscientiousness ($p = .004$). However, the mediation analysis showed that self-efficacy significantly mediated the relationship between these personality traits and risk-taking behaviors. The participants in Merritt and Tharp (2013) risked physical injury by performing challenging parkour maneuvers, with the potential reward being personal achievement, skill development, and social recognition within the parkour community. The runners overcame a primary potential cost (e.g., physical pain and injury) for opportunities with little extrinsic benefit. The opportunity is dependent on the player and their circumstances.

The 46 women who manufactured methamphetamines in Deitzer et al. (2021), for example, risked personal freedom (i.e., jail time) along with injury and even death. In their semi-structured interviews and the subsequent thematic analysis, the researchers found the same pattern of higher self-efficacy for "cooking" meth resulting in greater acceptance of risk. The women had clearly different levels of efficacy beliefs, and they all were involved in the same process. However, the women with higher self-beliefs focused, alongside the financial gain, on the respect and control that came with cooking—an intrinsic benefit in exchange for their skill. The women with low efficacy beliefs avoided cooking—even if they were still involved—experienced greater anxiety in the process, and their "rewards" were entirely extrinsic—money and a steady supply of drugs.

Success in measuring the unobservable self-efficacy is dependent on creating an environment in which players may demonstrate indicative behaviors while engaging in a task (Mislevy et al., 2003). The decision to engage in greater risk is dependent on the person's perceived balance of risk versus opportunity. Using risk-taking as an observable indicator of self-efficacy starts with establishing the connection between the two constructs, but certain influences on risky decisions are observable and others are not. For example, players set their own goals, and those goals are not always aligned with the goals provided by gameplay (Slater et al., 2017). Certain players are more driven by beating their high score, outperforming their peers, or finding all the hidden characters in an open game world. Players with high self-efficacy beliefs are expected to make riskier attempts

but may not be focused on accumulating extrinsic gains (e.g., coins, points, etc.). Games can exist in complex worlds that give players a diverse selection of incentives, opportunities, and rewards. Thus, games meant to measure self-efficacy using risk-taking should include multiple opportunities for risk with varying incentives.

The flexibility and appeal of games do come with a potential caveat related to risk-taking: games provide a place for players to engage in behaviors relatively free of real-life consequences. The risk tied to in-game incentives does not have the same impact that loss of freedom or money may present in real life. For example, Beghetto (2009) conducted a survey-based correlational study involving 585 elementary school students to examine the relationship between intellectual risk-taking and self-efficacy in science. The study used correlation and regression analysis to explore this connection. Correlational analysis showed that intellectual risk-taking (IRT) was positively correlated with self-efficacy in science ($r = .56$, $p < .05$). This suggests that students with higher self-efficacy are more likely to engage in intellectual risk-taking behaviors. Additionally, IRT had a weaker correlation with actual science ability ($r = .16$, $p < .05$). Hierarchical regression results indicated that self-efficacy was the strongest predictor of IRT ($\beta = .27$, $p < .01$), second only to interest in science ($\beta = .33$, $p < .01$). In this context, students risked making mistakes or facing criticism for their intellectual contributions, with the potential reward being increased knowledge and academic success. Beghetto (2009) also noted that students are more likely to take intellectual risk on game-like tasks than on school-like tasks. However, Beghetto (2009) used a self-report survey to measure IRT while also providing strong support that self-efficacy influences the decision to take a risk. The use of games could inflate risk-taking observations, or students who still take risks and seem high on other measures may be negatively affected by a self-efficacy scale similar to an exam (i.e., self-report). Thus, the concern is that the risk-taking behavior may be inflated compared to the goal-setting and persistence behaviors and the self-report.

Regardless, these works show a clear predictive relationship between self-efficacy and risk-taking. While risk-taking will require extra consideration in analysis, it remains a valuable behavior in measuring self-efficacy via gameplay behaviors.

3.5.3 Goal Setting

Setting a goal is a process of identifying specific attainable outcomes that gives organisms direction, a drive to persist, and a deep sense of accomplishment when achieved. In their seminal work, Locke and Latham (1991) outlined criteria for setting goals and the benefits that setting goals

can bring. They also demonstrated how the relationship between assigned goals and performance is mediated by the person's efficacy beliefs. Bandura (1997) also extensively reviewed the relationship between efficacy beliefs and goal setting. There is ample evidence that higher efficacy beliefs result in more challenging personal goals (Appelbaum & Hare, 1996; Bandura, 1997; Locke et al., 1984; Schunk, 1984; Zimmerman et al., 1992). In other words, high efficacy beliefs encourage individuals to set higher goals for themselves because they see the potential for greater success within the reach of their resources and capabilities. Those with weaker efficacy beliefs set lower goals that seem more "within reach" of their capabilities.

Relevant to the game design, challenging assigned goals—especially when assigned by a trusted source—likely "raises self-efficacy because this is an implicit expression of confidence by the [source]" (Locke & Latham, 2002, p. 709). A well-designed game, one that appeals to various playstyles, provides multiple mechanisms through which players can set their own goals (i.e., personal goals), which are more indicative of efficacy beliefs because assigned goals influence both self-set goals and self-efficacy beliefs. A game could give a solitary goal, which many do, while also providing multiple pathways to achieving the overarching goal. For example, the overall goal for *Mean Alchemy* is admission into one of the "guilds". However, the player can choose the quests or tasks that all provide progress toward the overarching, or distal, goal. This means the moment of observation must be on proximal goals—the steps players take to reach the distal goal. Focusing on proximal goals also sidesteps cases where a player does not accept the assigned proximal goal and instead plays the game to maximize a sub-score or master a particular mechanic.

Assessing Goal-setting as a Measure of Self-efficacy. Because of the extensive body of research establishing the connection between self-efficacy and goal setting, recent studies continuing to explore this relationship are relatively scarce. Two counterexamples are Zimmerman et al. (1992) and Wood and Bandura (1989), both of which are frequently cited articles.

Zimmerman et al. (1992) investigated the role of self-efficacy in setting goals for academic grades. They used surveys and direct observations to collect data on 102 high school students from two separate high schools. The authors used correlational analysis followed by path analysis to explore causal linkages between prior grades, efficacy beliefs, parents' assigned goals, students' self-determined goals, and students' final grades. Their correlational analysis showed strong correlations between efficacy beliefs and self-set goals ($r = .41$, $p < .05$). Students' self-set goals were the strongest causal link to their final grades (i.e., their performance; $\beta = .43$, $p < .05$). Parents' assigned grade goals had the same influence on students' self-set goals as students' efficacy beliefs

for academic achievement ($\beta = .36$, $p < .05$), but it is probable that the assigned goal given by a learning game would have less impact than a goal passed down from an authority figure like a parent or teacher. That point remains an assumption in need of further exploration. The takeaway from Zimmerman et al. (1992), beyond the empirical linkage between efficacy beliefs and goal setting, is that the context and source of the goal are important. Providing students with the agency to set their own goals may provide a better indicator of their efficacy beliefs than goals handed down from a "higher authority".

The need for autonomy is further demonstrated by the series of three studies conducted and synthesized by Wood and Bandura (1989) into a comprehensive path analysis. The two researchers explored multiple aspects of the managerial decision-making process and how it is influenced by efficacy beliefs, perceived controllability, task difficulty, and business goals. The group with greater controllability in the decision-making process (i.e., higher autonomy or influence on the situation) set increasingly higher goals over three trials compared to the low-controllability condition, which showed a steadily decreasing goal over the same three trials.

Most impactful, however, is the final path analysis across the three studies showing that self-efficacy had a significant and consistent causal influence on personal goals in both the first instance of observation ($\beta = .25$, $r = .41$, $p < .05$) and the second ($\beta = .62$, $r = .65$, $p < .05$). Interestingly, the strength of the influence of efficacy beliefs on personal goals increased between instances, with the correlation increasing from $r = .41$ to $r = .65$. The two correlations are separated by a completed managerial performance. The increasing influence of efficacy beliefs between performances also exemplifies the dynamic nature of self-efficacy. The beliefs one holds about the ability to accomplish a task are in continual flux, dependent on the environment, task difficulty, and perceived path to success. Completing a one-off survey ignores these complex conditions in an attempt to reduce efficacy beliefs to a single score. The benefit that a game-based assessment provides for measuring efficacy beliefs is the ability not only to simulate the environment and context—much like the managerial decisions in Wood and Bandura (1989) were simulated—but also to allow for continuous assessment of efficacy beliefs over time and across tasks and difficulties to capture the dynamic nature of the construct more fully. This is likely true for self-efficacy just as it would be for many unobservable self-belief constructs.

Including measures of self-belief constructs like self-efficacy would improve the "fairness and validity" of important life-influencing assessments (Kyllonen et al., 2005, p. 176). Self-belief constructs (e.g., motivation, effort, perseverance) have an impact on important life outcomes (e.g.,

academic achievement) and are, potentially, more important than demonstrated competencies and skills (Gutman & Schoon, 2013; Stankov & Lee, 2014). Self-efficacy is dynamic, potentially bolstered or undermined by simple comparison with a peer or a misinterpreted comment from a teacher. Being inherently subjective and challenging to measure, self-report and interviews have been the primary—if not the only—standard for measuring self-efficacy. Earlier attempts at creating more sophisticated measures (McQuiggan et al., 2008) struggled to overcome the bias in self-report for training data and the lack of theoretical guidance (Choi et al., 2023; Czerwiński et al., 2025).

The approach I am proposing is not novel, but it has been waiting until the technology became readily available to begin a potential paradigm shift in assessment design. The stealth assessment of self-efficacy—and potentially other self-belief constructs—could enhance the fairness, validity, and scalability of assessments, effectively benefiting educational practice and policy. The educational system in the U.S. has always been slow to adopt new technologies, but the time for making the leap to stealth assessment of self-beliefs has arrived.

CHAPTER 4

METHODS

My study focuses on the collection and analysis of trace data from a digital learning game I designed and created to teach introductory statistics concepts. The game, *Mean Alchemy*, is a single-player, turn-based strategy game in which players manipulate distributions of data to achieve specific goals. *Mean Alchemy* was designed to be engaging and educational, with mechanics that encourage players to experiment with statistical concepts such as mean, variance, and distribution shape.

I used the data collected from the game to learn which in-game behaviors (e.g., persistence on a task) may be indicative of players' statistical efficacy beliefs. The primary objective of my work is to explore which features (i.e., variables)—designed from theory to support detection of self-efficacy—may be useful for this purpose. The methods I have chosen are based on insights from the literature review, which underscores the value of unobtrusive assessment of self-efficacy—i.e., stealth assessment. Using ECD as a guiding framework, I designed the game and the data collection process to elicit behaviors that could serve as proxies for self-efficacy beliefs. The methods are therefore focused on the design of the game, the collection of trace data, and the analysis of that data to identify meaningful patterns related to self-efficacy. Therefore, ECD was the primary methodological framework guiding the design of the game and feature engineering (i.e., ECD provided the structure to establish theory-based connections between self-efficacy and the proxy behaviors), while the data analysis methods were chosen to evaluate the predictive validity of those features in relation to self-efficacy beliefs.

Where ECD began to fall short was in modeling the complex relationship between the behavioral features and the self-efficacy beliefs. ECD provides a strong framework for designing assessments and identifying potential indicators. A true assessment made with ECD would clearly identify the specific behaviors in a clearly mapped relationship upfront. However, due to the complexity of the self-efficacy beliefs, ECD alone was insufficient for capturing the nuances of these relationships. Therefore, I turned to EDM techniques and multiverse-style analysis to engineer many potential features and evaluate how well the engineered features could predict self-efficacy beliefs. This involved selecting appropriate algorithms, tuning model parameters, and validating the models using techniques such as cross-validation.

4.1 Research Design

In this study, I used a quantitative, nonexperimental predictive design grounded in educational data mining. No variables were manipulated. Instead, the study relied on pretest and posttest responses and gameplay telemetry to examine the extent to which behavioral proxies derived from gameplay could predict students' self-efficacy beliefs. The design is therefore observational and predictive rather than causal, with the central analytic goal being to evaluate how well theory-informed behavioral features support inference about self-efficacy in a game-based assessment context.

4.2 Participants

I designed the game with middle-school students in mind, when foundational statistics concepts such as mean, median, and mode are first introduced in formal coursework (see Section 4.4.1). However, I wanted to target a developmental period with the greatest chance of variability in statistical self-efficacy beliefs. Too early, and students may not have enough experience to have formed stable beliefs about their abilities. Too late, and students may have already developed strong beliefs that are less responsive to the game-based learning experience. Variability in efficacy beliefs is a key consideration in this study. My sample includes students from 8th to 12th grade, which allows for a range of developmental stages and experiences with statistics. And I include a content knowledge measure to control for differences in statistical understanding. The sample is predominantly male, which is a common demographic pattern in video game contexts—and this was a video game design classroom.

4.2.1 Sampling and Recruitment

I collaborated with a local K-12 research school in the Southeastern United States to enter their Video Game Design classroom for two days, aided by the instructor. Because the research school relies on a random lottery for admission, my sample's diversity in terms of socioeconomic status and academic background—although girls are underrepresented due to the male-dominated nature of video game contexts—was still representative of the local demographics, which supports the generalizability of the findings on those dimensions. The school administration facilitated communication with parents and students. The informational sheet emphasized the voluntary nature of participation and the educational value of the game-based learning experience.

The full gameplay telemetry data included 109 students. From that larger pool, 102 students had gameplay data matched with pretest survey records and therefore contributed the demographic and baseline score information reported here (Table 4.1). Posttest availability was smaller and more variable by measure. Dropped students were absent at least once during the study period. Two students only completed the content posttest and not the self-efficacy posttest. Only one student was dropped for not participating (i.e., sleeping). For reasons explained later, the pretest score was chosen as the target for prediction. The model results reported in Chapter 5 use a common analytic subset of 86 students so that the model comparisons, confusion matrices, and change analyses are all grounded in the same cases.

Table 4.1: Participant counts and gameplay exposure

Block	Measure	Value
Participant counts		
	Students with gameplay logs	109
	Students with matched pretest data	102
	Students with posttest content scores	95
	Students with posttest self-efficacy scores	93
	Students in common analytic modeling subset	86
Gameplay exposure		
	Total play time, minutes, M (SD)	48.10 (15.80)
	Total play time, minutes, median	48.50
	Logged events, M (SD)	2456.69 (1036.23)
	Battles, M (SD)	48.91 (26.04)

The analytic modeling subset of 86 students is similar to other game-based learning and stealth-assessment studies that combine behavioral trace data with motivational or self-belief measures (Shute & Rahimi, 2021; Syal & Nietfeld, 2020; Wang et al., 2022). In game-based learning studies, the duration of gameplay is a pertinent question. I secured at least two full class periods with a minimum of 20 minutes of gameplay each day. Students generated over two million rows of behavioral data while playing. That matters for the interpretation of the later models because the behavioral features are being drawn from a nontrivial amount of activity rather than a single brief interaction.

The participants' score profiles are mixed, which partially informed the decision to use the pretest as the prediction target (Table 4.2). On the content measure, there was an average decrease of -0.65 from pretest to posttest, but the higher variability hints that students may not have taken the posttest as seriously as the pretest. That does not prevent posttest modeling, but it does make

Table 4.2: Assessment and survey descriptives

Measure	n	M	SD
Content pretest score	102	6.60	1.88
Content posttest score	95	5.96	2.10
Pretest self-efficacy	102	57.07	20.20
Posttest self-efficacy	93	58.48	22.29
Motivation composite	102	11.69	9.70
Math/statistics interest	102	0.20	5.42
Video game interest	102	8.13	3.55
Motivation to participate	102	3.36	4.51

the posttest target less stable than the pretest target. The self-efficacy scores are more stable, with a mean increase of 0.60.

Among the 102 students with matched pretest surveys, the mean age was 16 years (SD = 1.56). Table 4.3 summarizes age group, grade level, and gender alongside subgroup means and standard deviations for the main pretest measures used later in the analysis. Race was collected as a multiple-response item. Within the pretest sample, 65.7% identified as White, 25.5% as Black or African American, 12.7% as Asian, 2.9% as American Indian or Alaska Native, and 6.9% as Other; 13.7% selected more than one race, and 4.9% identified as Hispanic or Latino.

Table 4.3: Participant demographics and pretest descriptive statistics for students with matched pretest survey data

Characteristic	n	%	Self-efficacy pretest M (SD)	Content pretest M (SD)	Math/stat interest M (SD)	Video game interest M (SD)	Motivation to participate M (SD)
Age group							
14–15	42	41.2	59.10 (20.78)	6.60 (1.90)	-0.17 (5.32)	8.29 (3.49)	4.21 (4.46)
16–17	43	42.2	56.13 (21.25)	6.42 (1.88)	0.23 (5.47)	7.58 (3.97)	2.65 (4.80)
18+	17	16.7	54.44 (16.24)	7.06 (1.89)	1.00 (5.79)	9.12 (2.23)	3.06 (3.67)
Grade level							
8th Grade	30	29.4	61.45 (18.94)	6.50 (1.93)	-0.30 (5.63)	8.23 (3.73)	4.40 (4.21)
9th Grade	20	19.6	52.02 (22.81)	6.50 (1.85)	-1.25 (5.44)	7.75 (3.70)	2.90 (5.35)
10th Grade	17	16.7	60.29 (18.66)	6.47 (2.07)	0.71 (4.22)	8.06 (4.29)	2.29 (4.45)
11th Grade	23	22.5	53.21 (22.55)	6.61 (1.70)	0.39 (5.94)	8.00 (3.32)	3.00 (4.76)
12th Grade	12	11.8	57.34 (15.26)	7.17 (2.12)	2.75 (5.17)	8.83 (2.37)	3.75 (3.28)
Gender							
Male	83	81.4	58.86 (18.48)	6.72 (1.85)	0.61 (5.32)	8.57 (3.18)	3.52 (4.54)
Female	18	17.6	51.96 (23.30)	5.94 (1.98)	-1.44 (5.72)	6.22 (4.57)	2.83 (4.49)

Note. Values in the measure columns are subgroup means with standard deviations in parentheses. The interest and motivation columns are centered composite sums of four 7-point items after reverse coding, so negative values indicate ratings below the scale midpoint. Categories with very small group sizes were omitted to protect participant confidentiality.

4.3 Ethics and Risk of Harm

Working with underage populations requires greater attention to consent, privacy, and data protection. The study was minimal risk. Students completed surveys and played a digital learning game during normal school hours. The primary ethical considerations were therefore informed consent, confidentiality, and the secure handling of survey and gameplay records.

The study materials explained the purpose of the research, the voluntary nature of participation, and the kinds of data being collected. Students were informed that they could skip survey items or stop participation without penalty. Detailed study information and "opt-out" documentation were provided to parents and guardians by the school administration. The school frequently facilitates research studies and uses an established parental opt-out structure for minimal-risk research conducted during the school day. Accordingly, parents and guardians were notified about the study and given the opportunity to opt their child out rather than being asked to sign a separate permission form for each individual study. My study was approved by the university's Institutional Review Board (IRB) under study number STUDY00006193. Copies of the approval correspondence, parent information sheet, and parental opt-out form appear in Appendix A, Section A.1, and Section A.2.

Protecting participants' privacy and confidentiality was a central concern. I wanted to maximize the richness of data while removing any need to collect direct identifiers. I gave slips of paper with unique codes to each participant. The student wrote his or her name on the paper and, once they logged in, returned the paper to their teacher. On the following day, the teacher distributed the codes again. Once the study was completed, we destroyed the slips of paper. This process allowed me to link the gameplay data to the survey responses without retaining direct identifiers in the analytic dataset. Because the school used its established parental opt-out process, I did not retain signed permission records or direct parent contact information.

Survey responses and gameplay logs were stored on secured systems, and only the minimum information needed for matching and analysis was retained (i.e., study codes). Reporting in this dissertation is limited to aggregated results, model summaries, and de-identified examples, which reduces the risk of exposing any individual student. By relying on ordinary gameplay activity and brief surveys rather than intrusive procedures, the design supports the ethical goal of studying self-efficacy in a way that is both informative and respectful of student well-being.

Figure 4.1: Study code entry screen used to link survey responses and gameplay records without retaining direct identifiers in the analytic dataset.

4.4 Instruments and Materials

4.4.1 *Mean Alchemy*

Mean Alchemy is a story-driven educational game that places players in the role of a resourceful orphan living in a town where social and economic disparities challenge their ability to achieve stability and self-worth. The player dreams of joining the prestigious Alchemist Guild, a goal that promises them a home, a sense of belonging, and a future. However, the path to membership in the guild is not easy. Most guild members secure their place through inherited privilege, wealth, or expensive training—none of which are accessible to the player.

The story begins when the player takes on an odd job for a reclusive elder who has distanced themselves from the guild. During this task, the elder's familiar, a magical creature, takes a surprising liking to the player. Intrigued by this connection, the elder begins asking questions about the player's motivations and circumstances. Recognizing untapped potential, the elder offers to train the player in alchemy, setting them on a journey to develop their skills, gain reputation with the local community, and ultimately earn a place in the guild through resilience and ingenuity.

Learning Objectives. The primary instructional objective of *Mean Alchemy* is to provide players with an engaging, interactive space to develop a deep and intuitive understanding of statistical concepts related to central tendency and spread. The player's quests act as opportunities to practice these concepts, ensuring that the gameplay is seamlessly integrated with learning objec-

tives. *Mean Alchemy* creates a playful yet meaningful platform for students to explore and master these ideas.

The game is designed to help players achieve specific learning objectives related to statistics. Through their interactions with the elder, the alchemy table, and their quests, players will learn to:

1. *match* a small dataset to its corresponding histogram.
2. *explain* how the mean and standard deviation relate to the shape and spread (s/s) of a given histogram.
3. *interpret* the mean and standard deviation of a dataset to *draw conclusions* about the shape and spread of a distribution.
4. *explain* how outliers affect the mean and standard deviation of a dataset.
5. *manipulate* the values of a dataset to produce a histogram with a given shape and spread.

Objective of Play. Good instruction starts with learning objectives, and good game design starts with a compelling gameplay objective. In *Mean Alchemy*, the gameplay objective is closely tied to the learning objectives, creating a cohesive experience that motivates players to engage with the statistical concepts. The player's overarching goal is to earn enough *reputation* points to join the Alchemist Guild, which serves as a powerful motivator for players to persist through challenges and develop their skills.

Reputation is built by completing quests, with each quest presenting a unique challenge tied to the player's mastery of statistical concepts. The quests are categorized into three tiers of difficulty—easy, medium, and hard—each offering corresponding levels of rewards. This tiered approach allows for incremental progression and accommodates players with varying levels of skill and experience.

The Cycle of Gameplay. *Mean Alchemy* unfolds in a series of structured phases, creating a repetitive cycle that reinforces learning while maintaining player engagement (Figure 4.2). The cyclical structure of gameplay ensures that players can engage in short, manageable sessions while maintaining meaningful learning outcomes. There is not a linear progression, which allows players to revisit and reinforce concepts as needed. Learning the content is not tied to completing or "defeating" the entire game. Players play until their skills surpass the difficulty of the game, which in theory should be associated with mastery either of the content, the gameplay mechanics, or both. This design reduces the need for extensive tutorials and sequential progress, allowing players to learn and improve naturally through meaningful rehearsal and exploration.

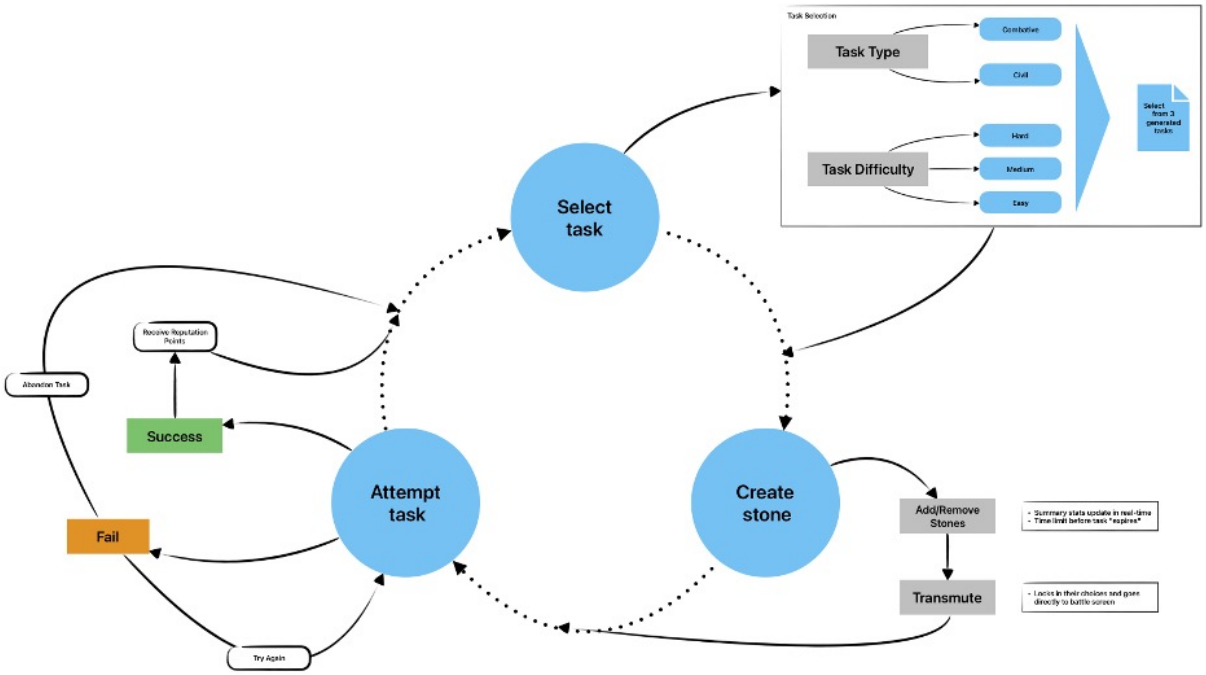

Figure 4.2: Cycle of gameplay in *Mean Alchemy*. The game loops through quest selection, stone construction, battle, and debriefing rather than progressing through a single linear sequence.

Players begin by selecting a quest from the bounty board, which displays available tasks with their difficulty levels and rewards. To help players strategize, the board provides information about the opponent's distribution, encouraging thoughtful planning (Figure 4.3). Players can choose to abandon a quest at any time, fostering a sense of safety and autonomy.

Figure 4.3: Selected quest view from the bounty board in *Mean Alchemy*. The board gives players the information needed to judge the level of challenge before committing to a battle.

Next, players move to the alchemy table, where they combine numbered stones representing raw materials to create a *Theta Stone* (see Figure 4.4). This stone embodies a distribution with specific statistical properties, such as mean and standard deviation. These properties directly affect the abilities of the player's familiar, including its attack strength, health points, and probability of dealing damage to opponents. Players can experiment with various combinations at the table, receiving immediate feedback as they refine their stone. Once satisfied, they proceed to the battle phase.

The battle screen presents a simple two-dimensional stage where the player's familiar faces off against an opponent's familiar (see Figure 4.5). The player can choose to Attack or Flee each turn, either dealing damage to the opponent or giving up. The battle continues until one combatant's health points are reduced to zero, with the winner gaining reputation. After each battle, players return to the lab to refine their strategies and prepare for the next quest.

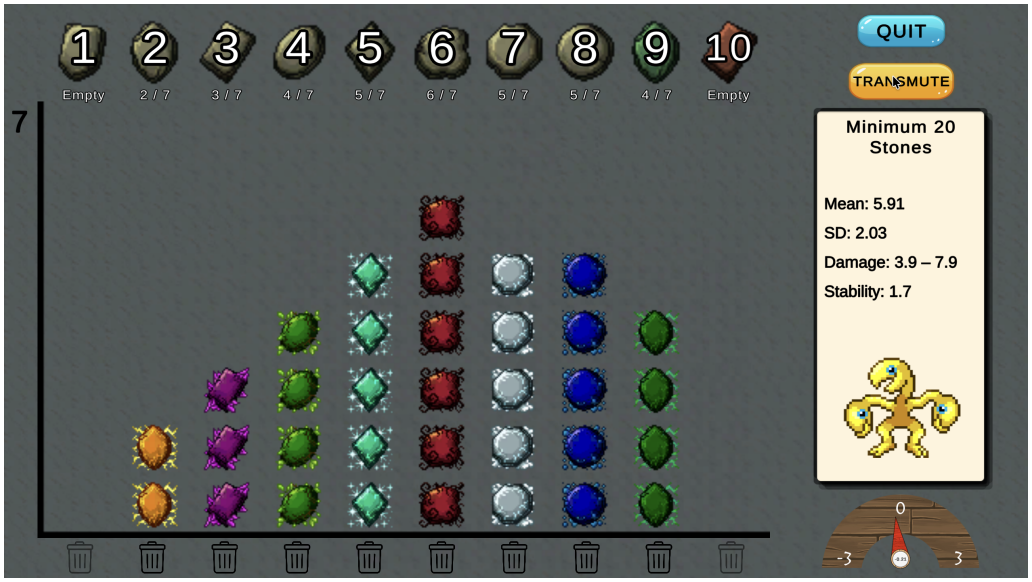

Figure 4.4: Alchemy table screen in *Mean Alchemy*. Players manipulate the numbered stones to construct a distribution that shapes the abilities of the familiar used in battle.

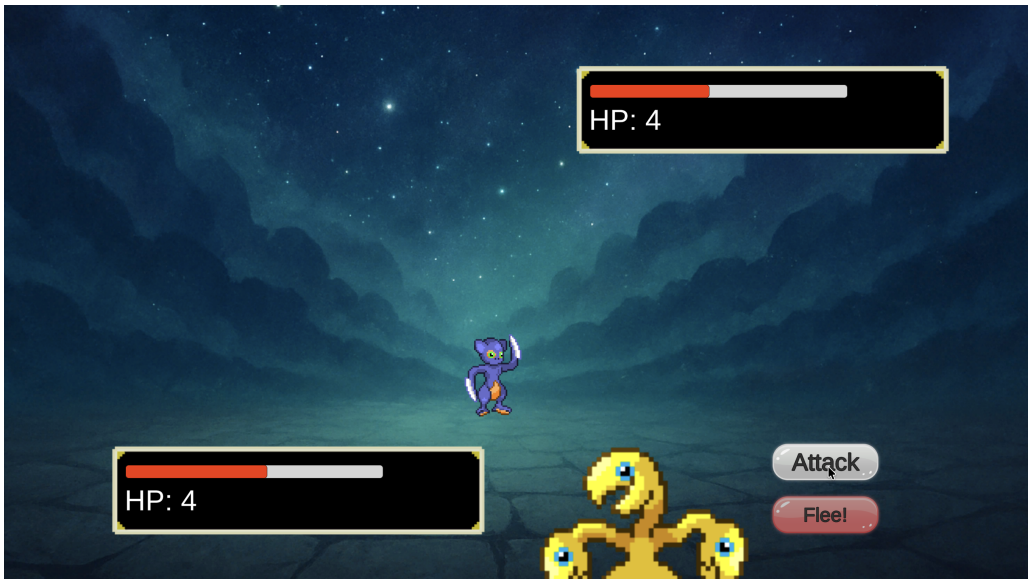

Figure 4.5: Battle screen in *Mean Alchemy*. The familiar's battle performance is determined by the distribution the player constructed at the alchemy table.

The familiar, a magical creature connected to the player's *Theta Stone*, serves as a visual and interactive representation of the player's statistical choices (Figure 4.4). Its abilities are directly influenced by the statistical properties of the stone, reinforcing the connection between abstract concepts and tangible outcomes. The tiered quest system offers players the opportunity to challenge themselves at their own pace, ensuring a balanced and rewarding experience.

Telemetry. Gameplay logs are the core behavioral evidence in this dissertation. The logging server records event-level actions such as quest selection, battle events, timing, and transitions between gameplay states. Functions embedded within the game send snippets of data to the logging server along with the user's unique identifier. Every action the player takes in the game is logged in this way. The telemetry data is unique in that there are levels of granularity unavailable in self-report data sources. The telemetry data, like adjusting the focus on a microscope, can cover raw action-by-action logs (e.g., clicking a button), sequences of behaviors (e.g., failing a task and then trying again), game states or series of values representing the states of complex in-game objects, and aggregated statistics summarizing overall gameplay (e.g., mean time spent on a task). The flexibility that comes with trace data allows for unique opportunities for planned and exploratory analyses.

The fine-grained telemetry data is then reduced into participant-level features that summarize behavior across play. Rather than starting with an existing game and searching retrospectively for usable signals, I designed *Mean Alchemy* and the candidate feature pool together with the specific goal of eliciting behaviors theoretically related to self-efficacy beliefs. This approach aligns more closely with ECD because the assessment opportunities were built into the game from the start rather than inferred later from whatever data happened to be available.

As the game, learning objectives, and target behaviors evolved over time, not all of the original ideas for observable indicators were feasible to implement in this version of the game. The final feature set therefore reflects both theory and design constraints. For example, risk-taking has a strong link to self-efficacy (Krueger & Dickson, 1994), so I designed the quest and battle systems to include meaningful reputational stakes. When a player failed in battle, they lost reputation points relative to the challenge difficulty. This reduced the value of brute-force play and created measurable opportunities to observe whether a student was willing to accept larger risks.

Feature Construction. The final model search used eleven candidate features representing persistence, risk-taking, goal-setting, and attitudinal context. These features were not intended

as universal indicators of self-efficacy. Instead, they were game-specific operationalizations of the broader proxy-variable logic used in this dissertation. Table 4.4 lists those modeled features by proxy family and provides a brief plain-language definition of each one.

Table 4.4: Modeled proxy features and brief definitions.

Proxy family	**Modeled feature**	**Brief definition**
Persistence	Failure event count	Total number of loss or failure events during gameplay.
	Planning episodes after failure	Number of return-to-planning episodes after a failed attempt.
	Total planning time after failure	Total time spent at the planning table after failures.
	Mean planning time after failure	Average time spent at the planning table after each failure episode.
	Total follow-up attempts after failure	Total number of attempts made after a failure event.
	Maximum follow-up attempts on a single quest	Largest number of immediate retry attempts observed on one quest.
Risk-taking	High-risk quests attempted	Count of higher-risk or higher-profile quests the student attempted.
	High-risk quest attempt ratio	Proportion of available higher-risk quests the student chose to attempt.
Goal-setting	Mean quest difficulty selected	Average difficulty level of the quests the student selected during play.
Bridge variables	Interest in math and statistics	Self-reported interest in math and statistics used to contextualize gameplay behavior.
	Interest in video games	Self-reported interest in video games used to contextualize gameplay behavior.

4.4.2 Self-efficacy Scale

The pretest and posttest phases include a statistical self-efficacy scale. Although Bandura (1997) explicitly outlined the proper specificity required to create a self-efficacy scale—i.e., written to target a certain domain and audience—there are few existing scales that *exactly* fit my needs. The best match I have found is from Carmichael and Hay (2009); their scale includes nine items that target middle school students' self-efficacy for statistical literacy. They obtained strong reliability (α

= 0.95) and a significant correlation with mathematical self-efficacy ($r = 0.56$, $p < .05$). Therefore, I used the nine-item scale from Carmichael and Hay (2009) and modified certain items that were not relevant to the covered content in my study (e.g., "show data correctly on a bar chart"). The full adapted instrument is included in Appendix B.

4.4.3 Demographic Questions

The demographic survey collected age, grade level, gender, race/ethnicity, video-game frequency, and related background variables. Additional questions asked about students' previous academic performance in mathematics, their access to technological resources at home, and experience playing video games, which could influence their engagement with the learning game. The survey was designed to ensure clarity and simplicity for a middle- or high-school audience. Collecting this data helped contextualize the study's findings, allowing for a deeper understanding of how different demographic factors may affect self-efficacy and learning outcomes in statistical concepts.

4.4.4 Content Knowledge

To measure content knowledge before and after the intervention, I created an assessment aligned with the statistical concepts introduced in *Mean Alchemy*. In other words, the content assessment evaluates students' understanding of measures of location, as specified in the Florida State Standards (MA.6.DP.1). The pretest serves as a baseline measure of students' initial knowledge, while the posttest allows me to assess their learning gains following gameplay. This tailored approach ensures that the content assessment is both age-appropriate and directly relevant to the instructional material covered in the study. The full pretest and posttest item prompts are included in Appendix C and Appendix D.

4.4.5 Motivation and Interest Questions

I also included a short motivation-and-interest battery at the start of the session. The items were organized into three small scales: interest in math/statistics, interest in video games, and motivation to participate in the study. Each scale included four items rated on a 7-point agreement scale, and each scale included reverse-coded items. After reverse coding, the four-item composites were centered on the scale midpoint before summing, so a value of zero indicates a neutral midpoint and negative values indicate lower endorsement. The full item sets are included in Appendix E. The item sets were drawn from the revised prospectus materials and served two purposes. First, they provided participant-level descriptors of engagement and value. Second, selected composites

were used as proxy variables in the predictive models so that persistence or challenge seeking would not be interpreted as self-efficacy in isolation (i.e., to control for value).

In the final modeled feature pool, two attitudinal proxy variables were retained: interest in math and statistics and interest in video games. The motivation-to-participate composite was preserved in the cleaned dataset and used descriptively, but it was not included in the final eleven candidate features used in the model search.

4.5 Procedure

The data collection was broken up into three phases: pretesting, gameplay, and posttesting. I was in the same classroom for two days for all but two periods—planning and an AP course that could not participate. On the first day, I introduced the study, distributed the pretest surveys, and helped students log in to the game. On the second day, I facilitated gameplay and then distributed the posttest surveys. The students were given a unique code to link their survey responses to their gameplay data without retaining direct identifiers in the analytic dataset (Figure 4.1).

I created a web-based study dashboard for the students (Figure 4.6). The links they needed (e.g., pretest, game) changed as needed per day. The page also included important directions, a walkthrough of the gameplay cycle, gameplay strategies, and a place to provide feedback on the game. The self-report instruments were hosted in Qualtrics and provided via the study dashboard, which the teacher provided via their classroom online portal.

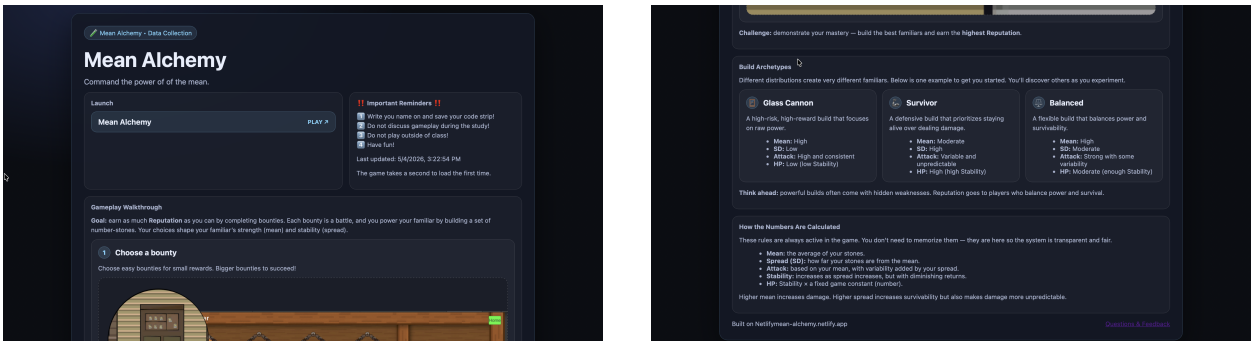

Figure 4.6: Study dashboard views used to guide participation. The dashboard centralized the links, directions, gameplay reminders, and support information needed during the two-day data-collection procedure.

Participants were discouraged from too much interaction (i.e., cooperative play), but cross-talk generated by excitement about gameplay was passively allowed. The classroom teacher and

I engaged only for the purposes of classroom management and technical support, but we did not provide hints or support for either content or gameplay directly.

The pretest included the self-efficacy scale, content assessment, and demographic questions. It took students about nine minutes to complete the pretest ($M = 8.98$, $SD = 2.12$). They began playing immediately after the pretest, so the pretest scores reflect students' beliefs and knowledge before any exposure to the game. The gameplay phase lasted for two class periods, with students playing for an average of 48 minutes across the two days. The posttest included the same self-efficacy and content measures as the pretest, but in alternative forms to reduce practice effects. The posttest took about five minutes to complete ($M = 4.57$, $SD = 1.68$). The posttest scores reflect students' beliefs and knowledge after their gameplay experience.

4.6 Data Analysis

I searched across a large space of reasonable choices for modeling the relationship between gameplay behavior and self-efficacy beliefs. The model search was designed to test whether a predictive signal for self-efficacy beliefs remained visible across multiple reasonable analytic choices rather than depending on one narrow modeling decision. Although this study does not implement a formal multiverse analysis, it follows a similar robustness logic by testing whether the signal persists across multiple defensible specifications (Steegen et al., 2016). The search included variations in feature (i.e., variable) selection, data preprocessing (e.g., normalization), model family (e.g., Naive Bayes, Logistic Regression), and hyperparameter configurations (i.e., specific adjustments made to individual models). The same feature pool and evaluation procedure were applied across both targets—pretest and posttest self-efficacy classifications—to ensure that the comparison of results was grounded in a consistent analytic framework. This approach aligns with the Results chapter's focus on model agreement because it allows an examination of how different modeling choices converge or diverge in their predictions of self-efficacy beliefs.

4.6.1 Outcome Preparation

The models included in the search all predict categorical outcome variables. Bandura (1997) stressed using a 100-point scale for self-efficacy ratings. I still discretized the self-efficacy scale into three groups: *high*, *medium*, and *low* self-efficacy. Doing so allowed me to use these specific models. However, more importantly than the model requirements, too much noise in the responses could make it difficult to predict a continuous score with precision. There is too much noise when it comes

to real-world assessment of self-efficacy. Rather, by looking at groups along a range—specifically, they were separated at the mean plus or minus one-half of the standard deviation—I could focus on distinguishing among meaningfully different levels of self-efficacy beliefs rather than predicting a continuous score alone. More simply, I'm interested in seeing how different the behaviors are between the higher and lower self-efficacy groups. I had to create a cutoff at some point. And the middle group is likely too noisy or overlapping to be helpful for interpretation. The middle group is still included in the modeling, but the main focus is on the high and low groups. The middle group serves as a buffer zone that allows for some uncertainty in the ratings without forcing a hard cutoff between high and low self-efficacy beliefs. This approach also aligns with the idea that self-efficacy beliefs are not always clear-cut and can exist on a continuum, so having a middle category acknowledges that some students may have moderate levels of confidence that are not easily classified as high or low.

4.6.2 Model Search Factors

Before fitting models, I ranked features based on their mutual information with the target outcome. The mutual information ranking is a screening heuristic that orders features by how strongly each one is associated with the target outcome on its own. In other words, it tests the information shared between a feature (e.g., quest difficulty) and an outcome (e.g., pretest self-efficacy group) without controlling for any other features. This step only helps prioritize features for the model search. The ranking allowed me to create lists from which the top-k and random subsets of features were drawn for testing.

I also tested multiple preprocessing conditions to determine if the signal was sensitive to scaling or dimensionality reduction. I tested models with the base features (i.e., no preprocessing), standardized features (i.e., z-scored), standardized features with PCA to 6 components, and discretized numeric features using 5 quantile bins. These variants were included because different model families can respond differently to data representation. For example, linear models may perform better with standardized features, while tree-based models may not require scaling but could benefit from discretization.

Four model families repeatedly emerged as reasonable candidates for this problem. To preserve interpretability, I included linear models: standard logistic regression and an L1-regularized variant, where the penalty can shrink some coefficients to zero and yield a sparser model (Rudin, 2019; Tibshirani, 1996). Naive Bayes served as a lightweight probabilistic baseline that often remains

useful in practice despite its strong independence assumptions (Hand & Yu, 2001). Gradient Boosting served as a more flexible ensemble method that can capture nonlinear structure through stagewise additive fitting (Friedman, 2001). By including these four model families, I could compare their performance and gain insights into which types of models were most effective for predicting self-efficacy beliefs based on gameplay behavior. However, each set of modeling choices came with a unique set of hyperparameters, or small adjustments made to model parameters that can have a large impact on model performance. The specific hyperparameters for each model are too numerous to define in complete detail here, but the key point is that I did not pick one arbitrary configuration for each model family.

Evaluation Procedure. All candidate configurations were evaluated under the same resampling procedure so model comparisons were based on a common test. All model runs used a stratified 20-fold cross-validation procedure with shuffling. This means that the data was split into 20 subsets (folds), and each fold was used as a test set while the remaining folds were used for training. Stratification ensured that the class distribution of the target variable (e.g., high, medium, low self-efficacy) was preserved across each fold, which is important for maintaining the representativeness of the training and test sets. Shuffling the data before splitting helped reduce any potential bias from the order of the data. Hyperparameters were not tuned in a separate inner cross-validation loop or evaluated on an independent held-out test set. Instead, each combination of model family, preprocessing, feature subset, and hyperparameter setting was treated as a candidate configuration and evaluated under this same 20-fold cross-validation procedure.

Model performance was evaluated with a profile of metrics because no single measure captured all of the behavior that mattered in this study. By viewing model performance across multiple metrics, I could make a more informed decision about which models to carry forward into the final results. Table 4.5 defines each of the metrics used for evaluation.

Metric	What it shows	Interpretation	Citation
Accuracy	Overall share of students placed into the correct group.	Higher values are better; a value of 1.00 would indicate perfect overall classification.	(Powers, 2020)
Macro-F1	How well the model stayed balanced across the lower, middle, and upper groups.	Higher values are better; stronger values indicate more even performance across the three self-efficacy groups.	
Precision	When the model predicted a group, how often that prediction was correct.	Higher values are better; stronger values indicate fewer false positive classifications.	
Recall	How often the model successfully recovered students who truly belonged in a group.	Higher values are better; stronger values indicate fewer missed cases within each group.	
Matthews correlation coefficient	A single summary of how well predicted and actual group labels lined up overall.	Higher values are better; values closer to 1 indicate stronger agreement, values near 0 indicate weak alignment.	
Cohen's kappa	Agreement between predicted and actual groups after accounting for chance.	Higher values are better; values closer to 1 indicate stronger agreement beyond chance expectations.	(Cohen, 1960)

Table 4.5: Evaluation metrics used in the modeling search.

CHAPTER 5

RESULTS

This chapter moves from broad to narrow in what the models showed; that is, I begin with the full search footprint (i.e., all models, hyperparameters, and features tested), then narrow to the three strongest pretest models, then to the final L1 logistic regression model, and finally to the confusion-matrix follow-up analyses. This reporting includes how much signal the models found, where the strongest agreement appeared, and what separated the cases that were easier to classify from the ones that were not. The central point is straightforward. I am not searching for a single best model. Instead, several reasonable models converged on the same general pattern, and that agreement in a faint signal found in a noisy environment matters more than any one algorithm in isolation.

5.1 Step 1: Full Search Footprint and Overall Model Performance

A total of 480 model configurations were evaluated across four model families. That large search footprint matters because it shows the signal was tested against genuinely different model assumptions rather than a narrow slice of variants. The figures below summarize the average performance for each target while also keeping the within-family spread visible. These are averages across all configurations in the archive, not the best run for each family.

5.1.1 Pretest Model Metric Profiles

For the pretest target, Naive Bayes had the strongest average profile, followed by standard Logistic Regression, then L1 Logistic Regression (Figure 5.1). Gradient Boosting was the weakest. The spread inside each family was real, but not chaotic: pretest accuracy varied most for the L1 model and least for Gradient Boosting, which matches the sense that some families were more sensitive to the feature choices than others. That is still the important part. The signal is not a one-off spike from a single lucky fit. It survives across the family averages.

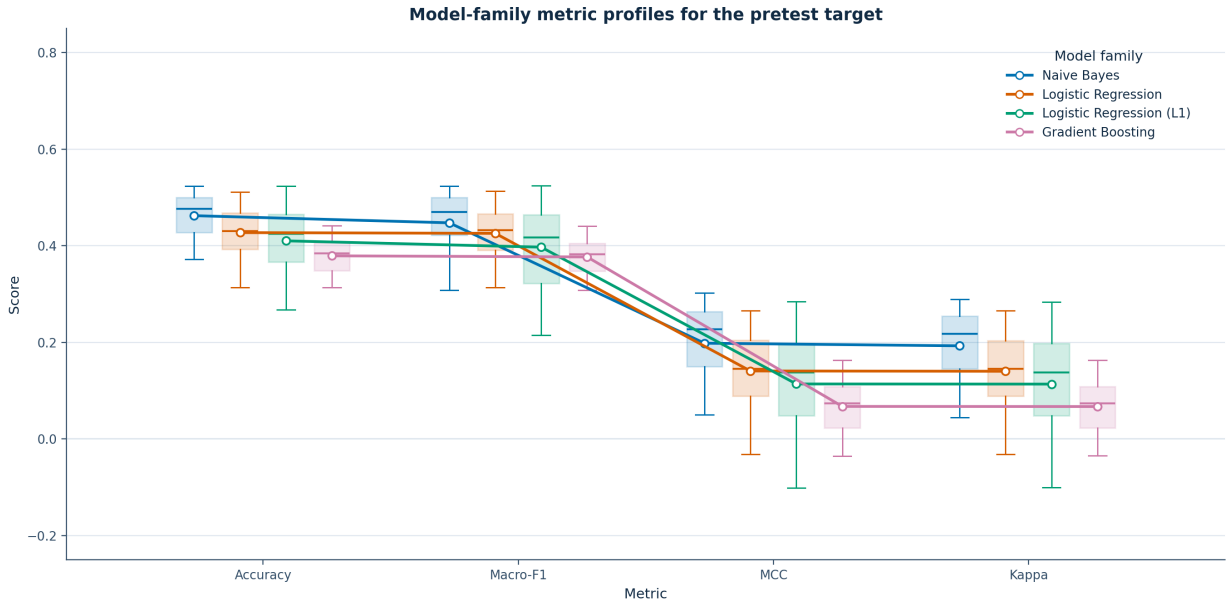

Figure 5.1: Model-family metric profiles for the pretest target across all evaluated model configurations. Each box summarizes the distribution within a family at a given metric, and the connected markers show the family means.

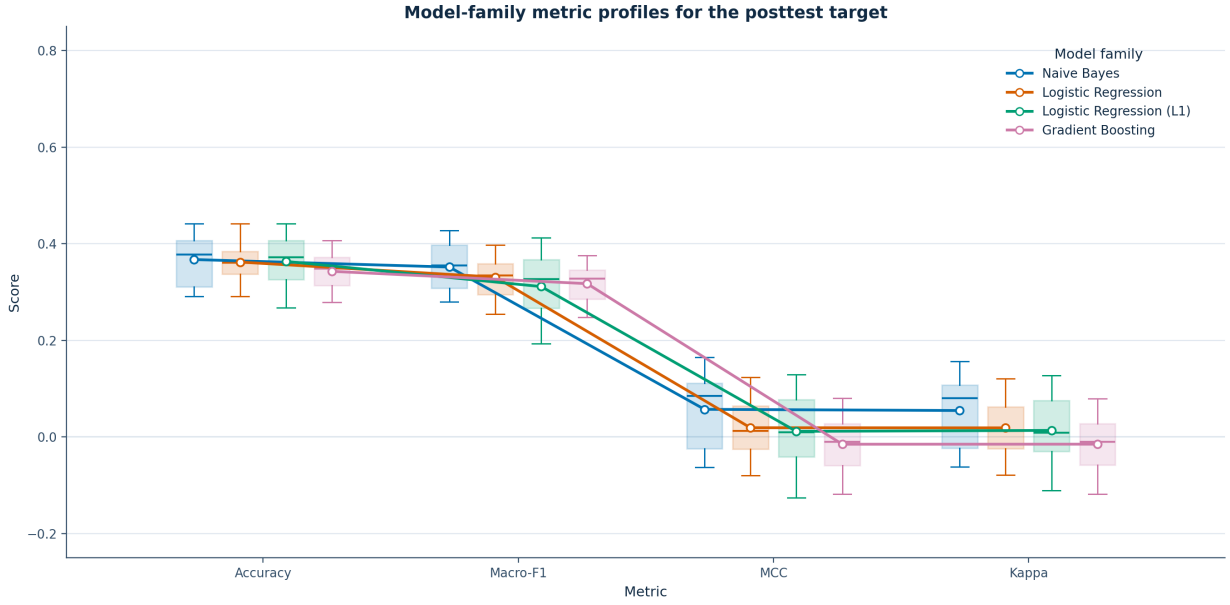

Figure 5.2: Model-family metric profiles for the posttest target across all evaluated model configurations. MCC and Kappa dip slightly below zero for some weaker configurations, so the shared scale extends below zero to preserve the full distribution.

5.1.2 Posttest Model Metric Profiles

Model performance for the posttest target is weaker across the board (Figure 5.2). Although the posttest target was weaker, that pattern is still informative. The same feature pool and the same model families are less stable once the outcome is moved to the later time point. The average scores compress, the differences between families narrow, and the whole picture feels flatter. That is useful in its own right, because it marks the posttest as the less predictable target and confirms why the pretest result deserves the main narrative. The family spreads are a little smaller around lower values here too, which reinforces the same point: the later target is not just weaker, it is less distinct.

Figure 5.3: Model comparison heat map for the pretest target

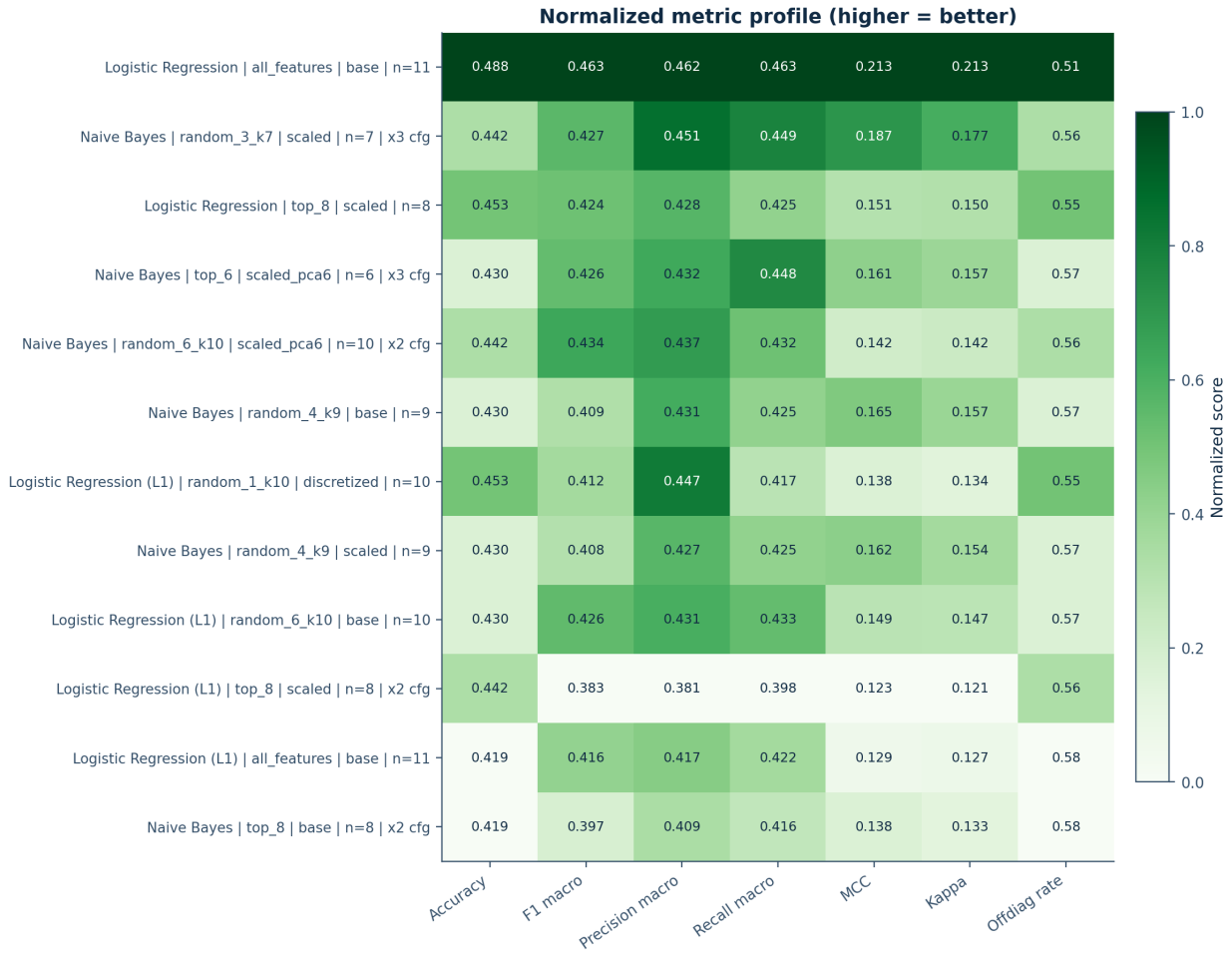

Figure 5.4: Model comparison heat map for the posttest target

5.2 Step 2: Agreement Among the Top Three Pretest Models

The heat maps (Figure 5.3 and Figure 5.4) compress the full search archive down to the metric profile alone. Each row represents one of the strongest retained model runs for that target. Cell color reflects a within-metric normalized score, so darker cells indicate stronger relative performance within that column, while the printed values remain the raw metric scores. The pretest target produces the clearer and more stable signal, so the remainder of this chapter stays centered on that target and treats the posttest target as secondary context. The heat maps also make clear that the three strongest pretest models do not diverge wildly from one another. They sit in the same region of the metric space, which supports a consistent signal amidst the noise.

Once the analysis is narrowed to the pretest self-efficacy groups, two model families are consistently at the top: L1 Logistic Regression and Naive Bayes. The L1 model has the highest accuracy and macro-F1. Naive Bayes is similar on precision and MCC. This interpretation does not depend on a single model specification. It survives contact with more than one reasonable model family. Their metric profiles are close enough that the story is not about a decisive winner. The same signal remained visible across different modeling assumptions. That agreement is the point of this section, and it is stronger than a single rank order would be.

5.2.1 Feature Importance and Comparison

The feature story is what matters most in this comparison. All three models lean on attempts after failure, mean seconds at table after failure, loss-event count, and interest in math/statistics. The two logistic models then keep the same more complete sets, which makes their agreement even harder to dismiss as random noise. Naive Bayes trims the feature set down further, but it does not move away from the same core signal. The fact that the same features are rising to the top across different model families is a stronger signal than any one model's feature importance alone.

Permutation importance is the companion diagnostic below because it asks a simple question: how much does model performance drop when one feature is shuffled while the others are left alone (Breiman, 2001; Fisher et al., 2019)? The permutation importance is an associational diagnostic of model reliance: it shows how much the fitted model depends on a feature for prediction, not the effect of intervening on the underlying data-generating process (Fisher et al., 2019). In Figures 5.5, 5.6, and 5.7, each bar shows the mean change in accuracy after permuting that feature, and the horizontal error bars show the standard deviation of that accuracy change across the repeated permutation runs. Larger positive values indicate greater model reliance, while values near

zero indicate that shuffling the feature had little stable effect on performance. These plots matter here because they allow the agreement story to move from "these models score similarly" to "they are leaning on the same kinds of variables."

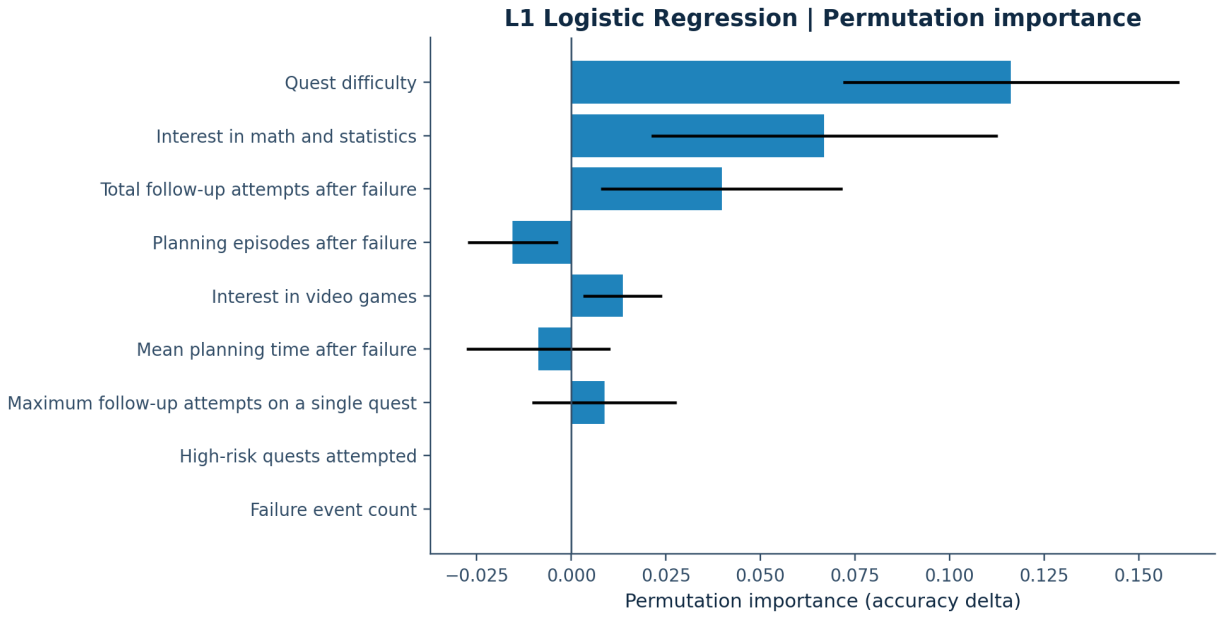

Figure 5.5: Permutation importance for the L1 Logistic Regression model

The feature-importance plots sharpen that agreement further. In the L1 and standard Logistic Regression models, mean quest difficulty and interest in math/statistics are top. But comparing the standard logistic regression model to the L1 model shows how the flattened features are washed out and the stronger features dominate. In Naive Bayes, interest in math/statistics and attempts after failure are most prominent. Across the three models, the same broader interpretation holds: challenge level, persistence after failure, and math/stat interest are doing most of the work. The exact ranking shifts a little, but the center of gravity does not.

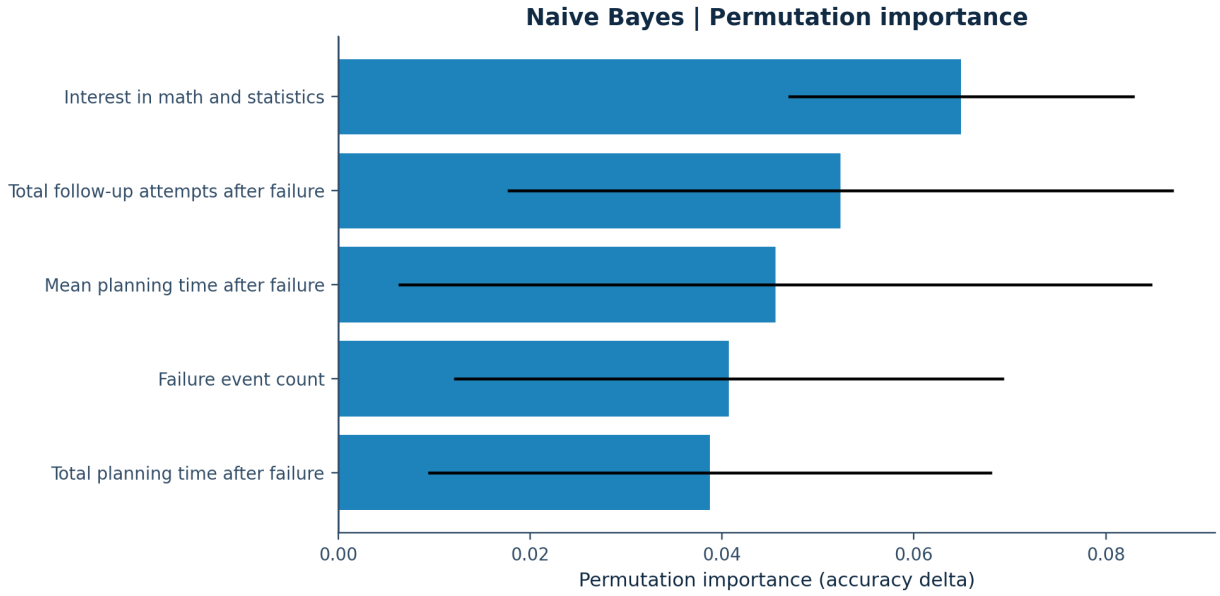

Figure 5.6: Permutation importance for the Naive Bayes model

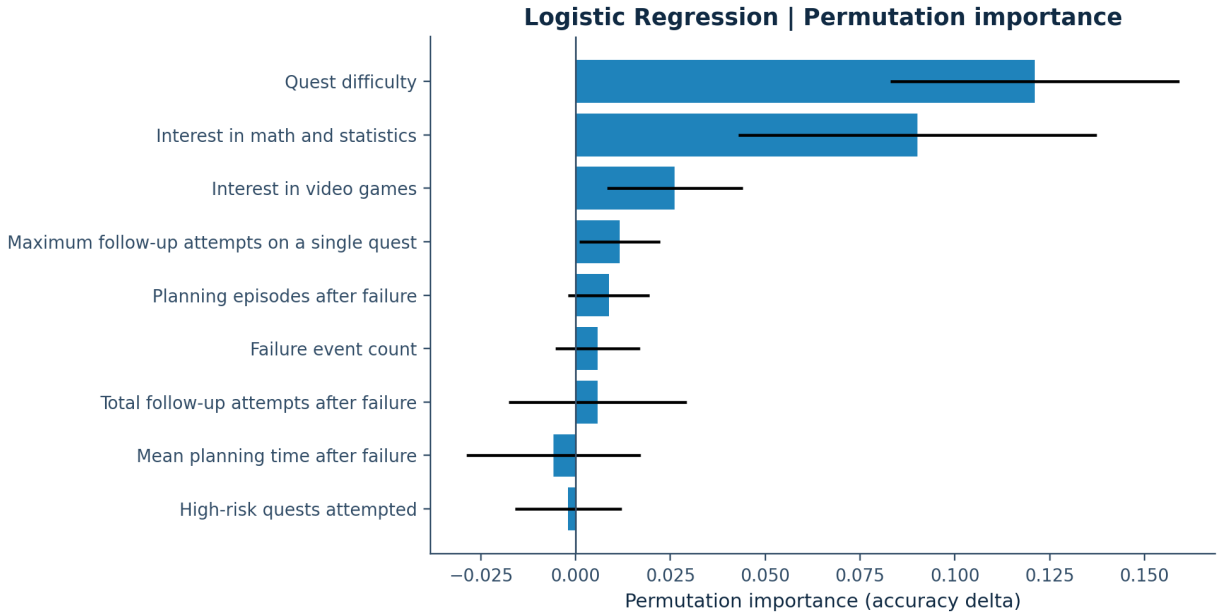

Figure 5.7: Permutation importance for the standard Logistic Regression model

5.3 Step 3: The Final L1 Model

I selected the L1 Logistic Regression model as the final interpretation model because it stays in the top performance band while also being easier to interpret. That matters here because I am not only trying to sort cases correctly; I am also trying to explain which behavioral traces are associated with those classifications.

L1 regularization is useful for that purpose because it suppresses weaker features rather than letting every candidate variable linger in the model with a small, noisy weight. In this fitted solution, that is exactly what happened. The selected subset contained nine candidate features, but not all nine carried the same interpretive weight, and two were effectively shrunk to zero. That makes the remaining pattern more disciplined. It also means the model is not just "the best" in a raw scoring sense; it is the cleanest summary of the feature story.

The coefficient column (Table 5.1) shows the signed effect inside the fitted logistic model: positive values push the prediction toward the upper self-efficacy class, while negative values would push it lower. Because this is a logistic regression model, those coefficients are reported in log-odds units, so a one-unit increase in a feature corresponds to a change in the model's log-odds of classification into the upper self-efficacy group, holding the other features constant. The SHAP-like values (Table 5.1) are global contribution scores, so they answer a different question: which features matter most overall when the fitted model is making predictions? They are not causal, and they are not interchangeable with coefficients. I am using both because the coefficient tells me direction, while the SHAP-like value tells me reach.

Three features stand out most clearly in the final model: mean quest difficulty, interest in math/statistics, and attempts after failure. Those are not the only features present, but they carry the clearest signal. Just as important, the weaker features do not disappear in an arbitrary way. The L1 penalty forces a decision. If a feature is not adding enough unique value, it gets pushed toward zero. That makes the surviving pattern easier to trust and easier to explain. It also keeps the final model from pretending that every weak signal matters equally.

Table 5.1: Feature interpretation for the final L1 Logistic Regression model

Feature	Upper coef.	Global SHAP-like	Interpretation
Quest difficulty	0.036	0.329	Strongest global signal; more informative for class boundaries overall than for the upper-class coefficient alone
Interest in math/stats.	0.513	0.289	Strong positive attitudinal signal for upper self-efficacy predictions
Attempts after failure	0.284	0.150	Persistence signal; students who kept trying after failure were more likely to align with higher self-efficacy
Attempts after failure per quest	0.000	0.101	Present in the candidate subset, but its direct effect was partly suppressed in the final solution
Mean secs. at table after failure	0.000	0.059	Small residual planning signal after regularization
Interest in video games	0.000	0.057	Present, but weaker than the math/stat interest signal
Planning episodes after failure	0.122	0.032	Small positive planning signal
High-risk quests attempted	0.000	0.000	Suppressed to zero in the final model
Loss-event count	0.000	0.000	Suppressed to zero in the final model

5.4 Following the Confusion Matrix

For the L1 model, the most important distinction is between on-diagonal cases, where lower and upper students were classified correctly, and off-diagonal cases, where those two ends were confused with one another (Figure 5.8). Table 5.2 reports the basic telemetry and age contrasts for those groups, and Figure 5.9 adds the scaled content and interest comparisons. Middle-class cases were left out of this comparison and grouped as *other*, because the main question here is what separated the clearest successes from the clearest misses. Furthermore, given the self-efficacy scores were discretized before analysis, the center values are inherently more ambiguous and less meaningful as a group than the clear lower and upper cases. The middle class is not unimportant, but it is not the focus of this particular follow-up.

Table 5.2: On-diagonal and off-diagonal comparisons for the L1 model

Measure	On-diag. mean (SD)	Off-diag. mean (SD)	Welch's t	p
n	32	10		
Gameplay sessions	2.156 (0.448)	2.100 (0.316)	0.441	.664
Gameplay events	2826.000 (938.111)	2674.000 (1174.714)	0.374	.715
Battle count	65.031 (27.961)	53.900 (23.005)	1.266	.222
Age	15.844 (1.370)	15.900 (1.287)	-0.119	.907

Several patterns are clear. Students on the diagonal had higher pretest content scores than students in the off-diagonal group (Figure 5.9). They also reported noticeably higher interest in video games and slightly higher interest in math/statistics (Figure 5.9). Posttest content scores, by contrast, were much closer across groups. Age was almost identical, which helps rule out a simple age-based explanation for the model's misses. The standard deviations matter here too: the off-diagonal group is more variable on the engagement measures, which fits the broader picture of a harder-to-pin-down group.

The telemetry variables point in the same general direction, though more cautiously. The on-diagonal group showed more activity, especially in battle count and total events, but the off-diagonal group is small enough that these comparisons should be read as directional rather than definitive. The practical point is that the clearest misses do not look random. They appear to cluster around weaker initial content performance and weaker engagement-related signals. This pattern is more informative than a simple misclassification reading would suggest.

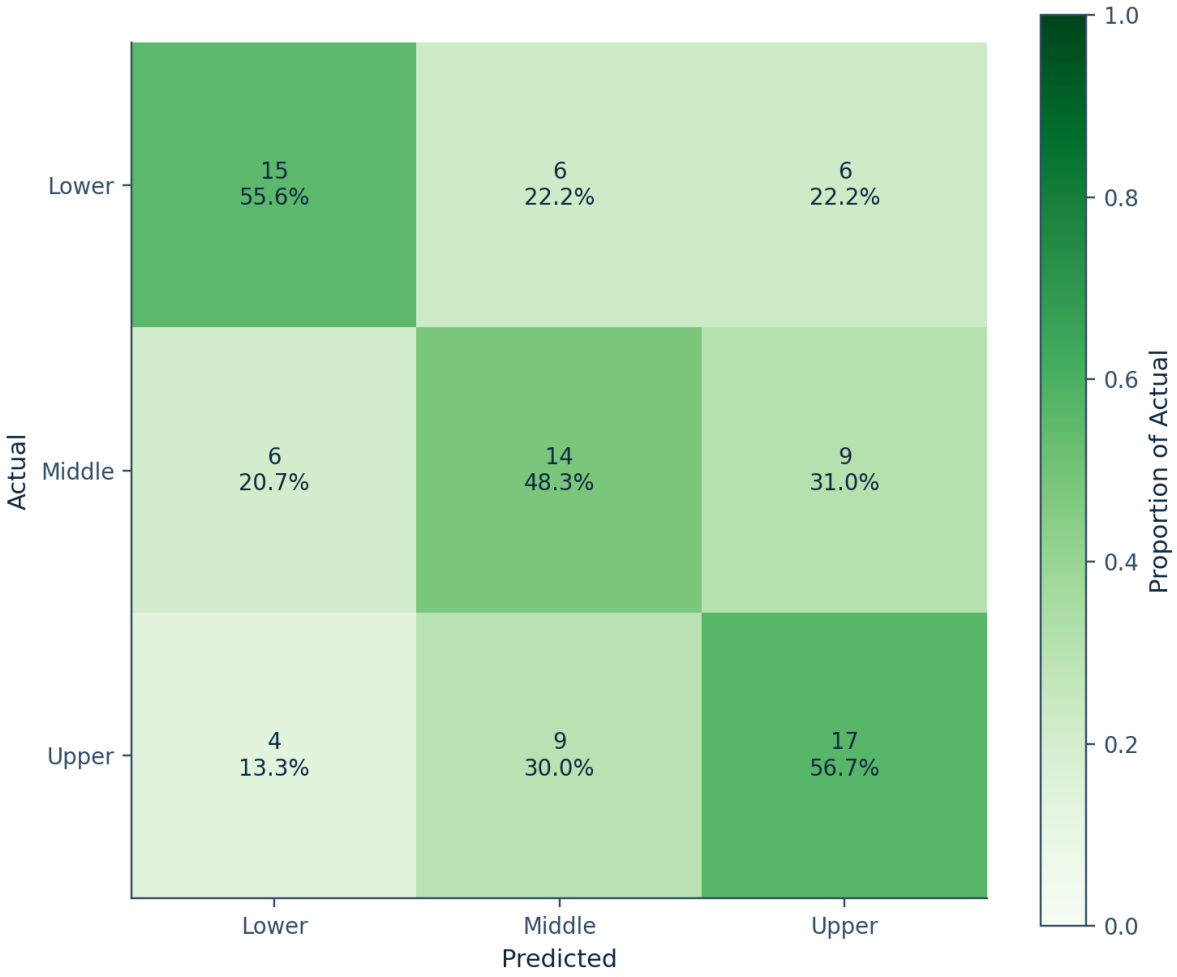

Figure 5.8: Final confusion matrix for the L1 Logistic Regression model

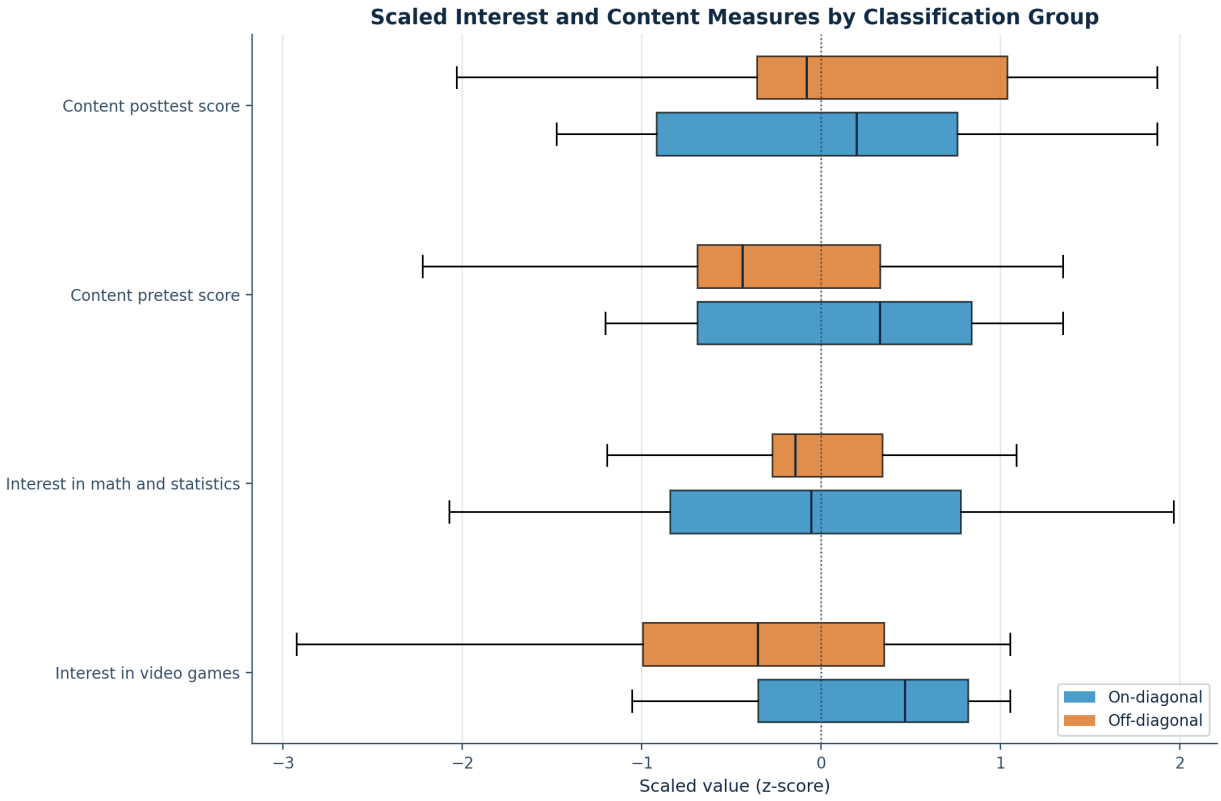

Figure 5.9: Scaled comparison of interest and content measures for on-diagonal and off-diagonal groups

5.5 Self-Efficacy Change Within Confusion Cells

The most interpretively important result emerged when I examined how self-efficacy beliefs changed from pretest to posttest. This is where the confusion matrix becomes more than a classification summary. It becomes a way to assess whether the hardest cases were also the most unstable ones.

For this analysis, the upper-left and lower-right cells are the correct endpoint classifications: upper-group students predicted as upper and lower-group students predicted as lower. The off-diagonal endpoint cells are upper-group students predicted as lower and lower-group students predicted as upper. Middle-class cases again fall outside this focused comparison.

Table 5.3: Self-efficacy change by confusion-matrix cell for the L1 model

Cell	n	Mean change	Mean absolute change	Increased	Decreased	No change
Correct upper	17	2.935	5.797	12	5	0
Correct lower	15	2.393	7.800	9	5	1
Upper predicted lower	4	-17.028	17.861	1	2	1
Lower predicted upper	6	1.222	14.407	3	3	0

Table 5.3 and Figure 5.10 show the pattern sharpening. Specifically, the off-diagonal cells show larger movement than the on-diagonal cells, especially when change is measured in absolute terms. Here the difference between the two change columns matters. Mean change is signed, so increases and decreases can cancel each other out. Mean absolute change removes the sign and therefore captures total movement regardless of direction. The most striking case is the upper-misclassified cell, where the mean change is strongly negative and the mean absolute change is the largest in the table. The lower-misclassified cell is different in direction, but not in instability. Its signed mean is close to zero only because the increases and decreases pull against each other; its mean absolute change is still large.

That distinction matters, because a small signed mean does not mean little movement. In the lower-misclassified cell, it means movement in both directions. Taken together, the off-diagonal cells suggest that the cases hardest to classify from a pretest snapshot were often the cases whose self-efficacy beliefs shifted the most afterward.

Figure 5.10 shows that the direction of change mattered as well. Correctly classified cases were more likely to show increases in self-efficacy from pretest to posttest, whereas cases involving decreased self-efficacy were more concentrated in the off-diagonal groups. In other words, the

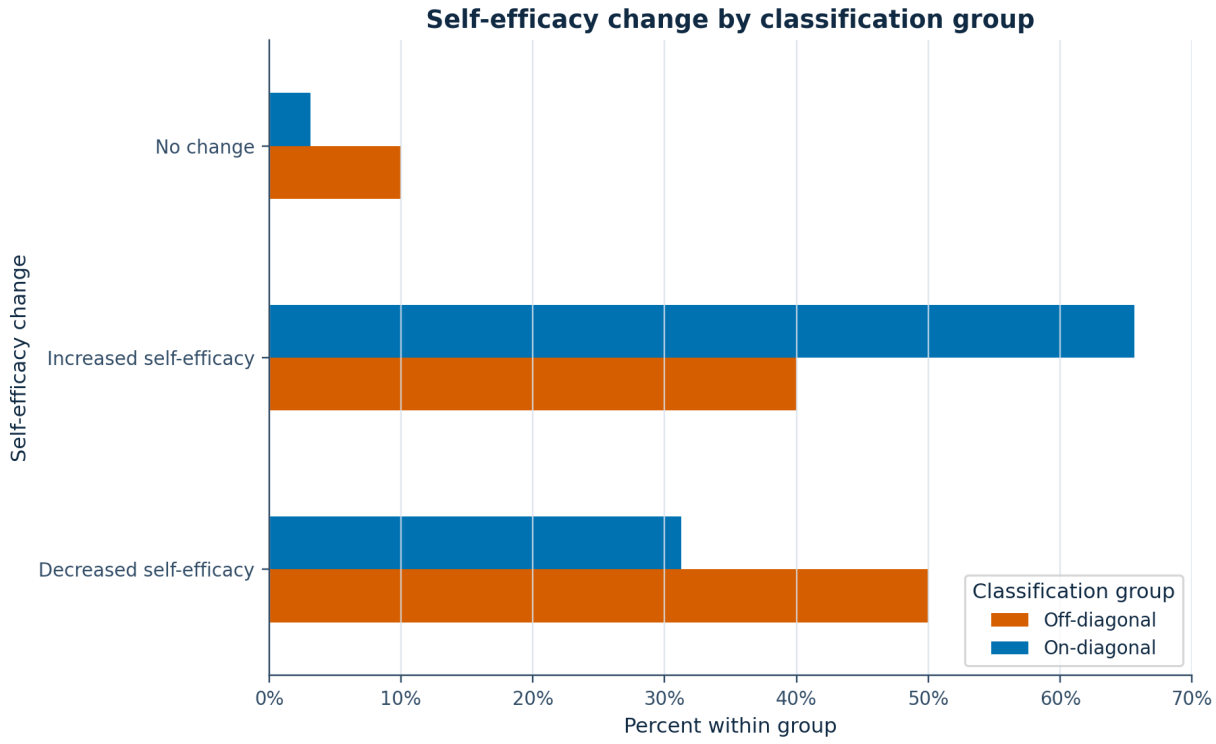

Figure 5.10: Distribution of self-efficacy change labels by classification group

hardest cases were not simply the ones with more movement overall, but often the ones in which that movement included a decline in reported efficacy beliefs.

We have consistent agreement between models, so the pattern may be reflecting the nature of self-efficacy. If beliefs are moving during or after the experience, then a static pretest target will be inherently harder to predict for those students. Cases with more stable self-efficacy beliefs may be more recoverable from a pretest-based model than cases in which those beliefs shift during the experience. Furthermore, a decrease in self-efficacy beliefs would be less likely without capturing evidence of a specific negative experience (e.g., series of failed attempts). That does not erase the signal in the model. It places a limit on what a one-shot classifier can reasonably capture.

5.6 Summary of Results

The results show a modest but real predictive signal for pretest self-efficacy group placement. That signal does not belong to one model alone. It appears across several model families, with the strongest agreement concentrated around challenge level, persistence after failure, and math/stat interest.

The L1 Logistic Regression model is the most useful final model not because it overwhelms the others, but because it keeps the same signal while making the feature story easier to read. Its confusion matrix follow-up also helps explain where the model struggled. The hardest cases were not simply noisier versions of the easy ones. They were more likely to involve lower pretest content, lower engagement-related signals, and larger shifts in self-efficacy between pre and post.

The current study supports the claim that behavioral proxies can carry information about self-efficacy beliefs. At the same time, it suggests that students whose beliefs are still moving may require a more temporal form of modeling than a single pretest snapshot can provide. The results are therefore not only about who was predicted correctly. They also show where a static snapshot starts to run out of room.

CHAPTER 6

DISCUSSION

The purpose of this study was to explore an evidence-centered assessment of students' statistical self-efficacy without relying on self-report, and to accomplish this purpose, I aimed to answer a single research question: **To what extent can behavioral proxies of self-efficacy in a game-based assessment predict self-efficacy beliefs?** And what I found does support there being promise in using behavioral proxies to assess self-efficacy. In response, I designed and developed a game with the explicit purpose, supported by decades of research, of eliciting behaviors indicative of players' statistical self-efficacy beliefs. Using telemetry from individual gameplay, I tested 480 iterations of four different statistical models to predict players' statistical self-efficacy before and after gameplay. In a statistically noisy environment with a mechanically simple game, I was able to detect a modest but consistent predictive signal. The contrast between the pretest and posttest metric profiles (Figures 5.1 and 5.2) makes that point visible at a broad level before the discussion narrows to individual models and features. Working toward the best predictive model (i.e., the highest metrics) is essentially searching for outliers. Rather, the consistent agreement among models is the strongest argument that a detectable signal exists. That emphasis on convergence across multiple defensible specifications follows the same general robustness logic used in multiverse analysis, even though the present study does not implement that method formally (Steegen et al., 2016). The strongest features among the models tested vary, but the ones that stay afloat are related to players' efficacy beliefs about statistics.

That being said, the signal is modest, and there is still a lot of noise in the data. The models are not yet at a point where they could be used for high-stakes decisions, but they do show that there is a signal to be detected. With further refinement of the game design, feature engineering, and modeling techniques, it is likely that the predictive power of these models could be improved.

The goal of this study was not to show evidence that any one in-game feature can detect self-efficacy. The process of translating theory into observable behavior (i.e., ECD) and then using *game telemetry* to unobtrusively measure those behaviors without intervention or the need for human scoring (i.e., EDM) is not new, but still relatively underdeveloped (McIntyre, 2023; McQuiggan et al., 2008; Shute et al., 2021; Ventura & Shute, 2013). My real contribution in this study was to

show that, when attempting to measure the contents of the *black* box, we can use modern statistical models to decode the complex patterns of "observable" behaviors that manifest as a direct result of the "unobservable". In this case, I showed that features derived to measure persistence, risk-taking, and goal setting, in combination, can reveal the signal of self-efficacy.

In one sense, this is the same foundation that underlies modern assessment more broadly: assessment has always depended on drawing inferences about what cannot be observed directly from what can be observed (Mislevy, 2003). The difference is that the remaining "hard-to-measure" constructs, such as self-efficacy, have largely remained dependent on self-report because the behavioral patterns were too complex to isolate consistently by hand. That dependence also matters because self-report measures do not always function equivalently across groups, which can create comparability problems in addition to ordinary response noise (Czerwiński et al., 2025). What changes here is not the logic of assessment, but the technological conditions under which that logic can be applied. With larger stores of trace data and statistical models capable of detecting complex behavioral structure, it is now possible to begin decoding patterns that have likely always existed in behavior but were beyond practical human interpretation. In that sense, assessment has been limited, in part, by technology. Researchers have had to wait for the infrastructure to collect this level of fine-grained data and for the computational power to make these kinds of models feasible at scale. Thus, the work at hand is to identify the combination of game mechanics and telemetry (e.g., ECD), feature engineering (e.g., EDM), and existing theory needed to make these hard-to-measure constructs interpretable through behavior.

6.1 Proxy Variables and Feature Agreement

No single feature can be assumed to serve as a universally useful predictor of self-efficacy beliefs across contexts. While it is possible to retroactively mine data to detect similar signals, if the environment (e.g., a learning game) does not elicit useful behaviors—ones that can be directly or theoretically connected to the target construct—there will not be a signal to detect. Furthermore, well-defined features in one environment (e.g., one game) will not automatically be useful in another game simply because of design differences between two different games—the game must be designed to optimize the opportunities for players to demonstrate the behaviors needed as evidence. The features from my study, however, can be generalized to guide future designs. We can use metrics like permutation importance and coefficients to determine which features (i.e., behaviors) matter and then reengineer similar features in different environments.

6.1.1 Feature Importance and Interpretation

There are two interest variables (i.e., interest in math and interest in games) that are still self-report. I considered these "bridge" variables, and they could eventually be engineered as behavioral features or incorporated into a more general student profile. Thus, they do not undermine the broader goal of inferring self-efficacy primarily from gameplay behavior. Fully engineering them was simply out of scope for this study. The self-report features also help control for, or factor out, the impact of intrinsic interest in math or video games that may muddy the signal for self-efficacy.

In the Naive Bayes model (Figure 5.6), interest in math and statistics dominates, but not by much. All features had positive permutation importance. That means the model suffered when any of the features were removed. There is a clear commonality in the importance of post-failure persistence features across all three models. However, the Naive Bayes model is dependent on the opportunities for failure and the actions immediately following. This emphasizes the fact that self-efficacy most commonly influences behavior in moments of success or failure. Bandura described efficacy beliefs as forward-looking judgments about capability (Bandura & Adams, 1977). That is, the beliefs influence the actions a person takes right after feedback. Thus, having the chance to fail and recover in the assessment environment (e.g., the game) is essential for detecting self-efficacy.

As it relates to the design of the game, the presence of failure is also essential. If a game is too easy or does not include the opportunity to fail or to "lose", then there is never an opportunity to *persist after failure*. That would make it impossible to detect the signal of self-efficacy beliefs. The Naive Bayes model—although interest is the top feature—is reliant on persistence variables to the extent that this model may be too focused on the persistence construct. However, looking at the model metrics, the Naive Bayes model has the strongest kappa, which means it is doing the best job of classifying players into the correct self-efficacy group (Figure 5.1). That suggests that persistence after failure is a strong behavioral proxy for self-efficacy beliefs in this context, and that the game design, similar to prior work on persistence, successfully elicited that behavior in a way that can be detected by the model (DiCerbo, 2014; Ventura & Shute, 2013; Ventura et al., n.d.).

Persistence importance in predicting self-efficacy is consistent with the theory (Wu et al., 2020). When a person experiences failure, their belief in their own efficacy is challenged. If they persist and eventually succeed, that can reinforce or even boost their self-efficacy beliefs. Conversely, if they give up after failure, that can lead to a decrease in self-efficacy (Bandura, 1997). The propensity of developing efficacy beliefs to fluctuate with experience is one of the reasons why self-efficacy is

difficult to capture with a single snapshot. Thus, patterns in behavior that occur over time (e.g., persistence after failure) are likely to be more informative than static features (e.g., total time on task) when trying to infer self-efficacy beliefs (Choi et al., 2023).

In the unregularized logistic regression (Figure 5.7), many features are retained but have a small impact when swapped out of the model. Still, the post-failure behaviors play a major role, but here the difficulty of the chosen quests has the stronger impact. The tiny positive permutation importance may hint at a predictive signal, but the overall contribution of these features is weak. Thus, the same model with L1 regularization applied helps wash away the noise and lets the stronger features stand out.

Logistic regression with L1 regularization was dominated by the difficulty of the quests chosen. Although weaker, the post-failure features remain present. Interestingly, the typical quest difficulty chosen had a weak upper coefficient of 0.036, but the strongest global SHAP-like value (Table 5.1). That means that as quest difficulty rises, it does not necessarily follow that the player will fall in the higher self-efficacy group. However, quest difficulty does have the strongest influence on which group the player is placed in (i.e., high, medium, or low self-efficacy), based on the highest SHAP-like value in the model. This result may indicate that the relationship between goal setting and self-efficacy is not purely linear—that is, we must look at the network of constructs (e.g., risk-taking, persistence, and goal setting) in relation to one another. Theoretically, as the person's self-efficacy beliefs rise, they would set and accomplish higher goals for themselves (Locke & Latham, 1991; Zimmerman et al., 1992). However, in the context of a learning game, there are similar behaviors that may not be directly related to the player's self-efficacy beliefs. For example, a player may choose more difficult quests in a strategy to maximize points or to finish the game faster, even if they do not have high self-efficacy beliefs (Baker et al., 2008). Thus, the game must be designed in a way that prevents clever workarounds to succeed without the relevant beliefs. In this case, the game design may have been successful in eliciting goal-setting behavior that is more closely tied to self-efficacy beliefs, but the relationship is still complex and may not be purely linear. Liu and Israel (2022) demonstrated the detection of problem-solving stages in students' gameplay telemetry from a popular puzzle game using a Hidden Markov Model. A similar approach may be more appropriate with goal-setting behaviors to capture patterns of behavior as the player moves closer or farther from their chosen goal (i.e., quest).

Alternatively, we would expect a stronger linear relationship between self-reported interest in mathematics and statistics and players' self-efficacy for statistics (Bandura, 1997; Beghetto, 2009;

Zimmerman, 2000). In the L1 model, a one-point increase in self-reported interest in mathematics and statistics corresponds to a positive increase in the log-odds of classification into the upper self-efficacy group (Table 5.1). Attempts after failure (i.e., persistence) are also well associated with self-efficacy (Bandura, 1997). Having interest and persistence clearly represented supports that the signal I am detecting is or is adjacent to self-efficacy beliefs. The choice of quests is also, as expected, globally influential. However, its direction is less clear. That could easily be explained by players' strategies, goals, and values.

Across the sampling space, persistence (i.e., time on task) is straightforward to implement. So, there were many more variables that fell under "persistence". Goal setting, even with just one feature, has the strongest impact. This also surprised me because the difficulty was not much more than labels displayed above the sets of quests and mild variations in quest parameters (Figure 4.3). I expected risk-taking to have a larger impact, but it is more likely related to the cost of the risk within the game. Real risk-taking requires the player to have a reasonable chance of losing something of value (e.g., money, points, or lives). In the context of a two-day study, the risk-taking variable either did not appear or was silenced by the L1 regularization. This is likely an indication that the way risk-taking was designed into the game was insufficient. Losing reputation points in a game the students will not play again does not have a meaningful negative impact on their lives. An authentic setting, such as performance on an important assessment, does inherently create an atmosphere of risk. For this reason, risk-taking is likely the hardest of the three constructs to observe, not because it is difficult to engineer the features, but because it is difficult to construct an environment of sufficient and reasonable risk.

6.2 Change in Self-Efficacy and the Need for Finer-Grained Modeling

Self-efficacy likely fluctuates at a rate faster than what can be aggregated from an entire session of gameplay. A person's belief in their own efficacy is always forward-focused, fixed on the task ahead. The decision to act hinges on whether that person believes their effort will produce the desired result or end in failure. The facts of the task are filtered through prior experience, and that combination of circumstance and memory produces the belief in one's efficacy (Bandura, 1997; Bandura & Adams, 1977; Zimmerman, 2000).

An apt analogy is with the combustion cycle in some aircraft, where the pilot still has to control the fuel-air mixture by hand. Prior experience is the fuel, the task ahead is the air, and self-efficacy

is the ignition that allows the engine to turn over and keep firing. If there is too little fuel—too little mastery experience—the engine does not start. If the mixture is off in the other direction, it does not start cleanly either. More importantly, the mixture can fail in the air just as easily as it can fail on the ground. A person may begin a task, but as difficulty rises or feedback accumulates, the balance between prior experience and the demands of the task can shift enough to interrupt continued action. That is part of why self-efficacy is difficult to capture with a single aggregated summary: it matters at the start of performance, but it also matters while the person is already in motion.

6.2.1 Pretest v. Posttest as Target?

The plane metaphor helps explain one of the clearest patterns in the results: pretest self-efficacy was more predictable than posttest self-efficacy. The same feature pool and the same model families that produced a modest but consistent signal for the pretest target became weaker and flatter at posttest, as shown in Figure 5.1 and Figure 5.2. By the time posttest self-efficacy is measured, the beliefs themselves may have shifted enough that the earlier aggregated behavioral signal no longer lines up as cleanly. In other words, some experience in the game may have influenced players' efficacy beliefs and they either increased or decreased. When working with aggregate features alone, such as I have done here, those changes are collapsed into an average of the entire experience—as has been done in most studies (DiCerbo, 2014; McQuiggan et al., 2008; Ventura & Shute, 2013).

The comparison between on-diagonal and off-diagonal cases points in the same direction. There were no meaningful differences in gameplay exposure between the on-diagonal and off-diagonal groups (Table 5.2), and the content and interest measures followed the same general pattern (Figure 5.9). At the same time, the off-diagonal cases tended to have greater variability in their content and interest in video games, but narrower variability in interest in math and statistics. This result could draw concern that what is being measured is as much pure interest as it is self-efficacy beliefs. However, the contrast with interest in video games suggests that the model is capturing more than general game interest and that the results are not driven solely by interest in video games.

Variability seems to hinder the predictive results to an extent in this study. The results in Table 5.3 and Figure 5.10 suggest that the hardest cases were often the ones in which players' self-efficacy beliefs moved the most from pretest to posttest, especially when those beliefs moved downward. One possible interpretation is that some players entered the game with relatively high confidence, made early progress, and then encountered difficulty that altered their self-appraisal

by posttest (Bandura & Locke, 2003; Kruger & Dunning, 1999). Under those conditions, a model built from aggregated behavior has less stable ground on which to classify the final belief state.

That matters because it places a limit on what an aggregated feature can reasonably do. A participant-level summary can still detect a meaningful signal, and the pretest results show that clearly. But once the construct itself is shifting during the experience, a single aggregate snapshot becomes a weaker representation of what the model is trying to predict. For example, if a player had consistent success up to the second-to-last stage, then their behavior up to that point (i.e., most of their gameplay) depicts rising efficacy beliefs. But that single final negative experience may be enough to cause the player to report low self-efficacy beliefs regardless of the amount of positive experience behind that single failure. In that sense, the weaker posttest signal is not just a lower-performing target. It is evidence that self-efficacy beliefs are too dynamic and therefore less detectable from aggregated telemetry. The practical implication is that future work should move toward finer-grained data, longer-form traces, and models built to recognize patterns across time rather than relying only on participant-level aggregates.

6.3 Conclusion

Developing a learning game is one path toward collecting useful behavioral evidence, but it is not the only one. Schools already collect rich data through grades, learning management systems, attendance, and other routine educational records. Widely adopted educational—or even purely entertainment-oriented—games may also contain useful behavioral traces. Data privacy concerns are valid in all of these settings and must be upheld to the highest feasible standard. At the same time, those opportunities only matter if data security is balanced with responsible educational use.

The benefit of a tailor-made research game is that it creates a more controlled environment. When a game is designed to elicit precise behaviors during the moment of gameplay—while novelty is still present—researchers can better isolate the signal of specific constructs and then use that knowledge to engineer features in existing data. That is one of the main contributions of this dissertation. The value of *Mean Alchemy* is not only that it produced a modest predictive signal for self-efficacy, but that it showed how game design, telemetry, and theory can be aligned from the start to make that signal observable.

6.3.1 Limitations

Using the self-report measures to train the model to predict self-efficacy beliefs is a limitation, but it is also a necessary step in the process. The goal of this study was to show that there is a signal in behavior that can be used to infer self-efficacy beliefs, and the only way to demonstrate that was to use self-report as the target variable. The next step is to build models that can predict self-efficacy beliefs without relying on self-report, which will require more complex modeling techniques and larger datasets. However, the current study provides a proof of concept that such models are possible and that there is a signal in behavior that can be detected.

The models were also built on a modest sample, which limits how strongly the results can be generalized. The signal was consistent enough across models to support the central claim of the dissertation, but a smaller sample also makes the estimates less stable than they would be in a substantially larger deployment. That matters most for the more complex patterns discussed in the posttest analyses and would matter even more for future temporal models that attempt to detect self-efficacy shifts across gameplay sequences rather than from participant-level summaries alone. A larger sample would also make it easier to estimate differences between the on- and off-diagonal groups and to evaluate the size of the pretest-posttest gap with greater confidence. More diverse sampling across different schools, grade levels, and game environments would help as well. The current sample was sufficient for a first-pass test of the central claim, but future work should prioritize larger and more diverse samples to test the stability and generalizability of the results.

The study is also tied to one game context. *Mean Alchemy* was designed specifically to elicit behaviors that could serve as proxies for statistical self-efficacy, and that design choice is one of the strengths of the study. At the same time, it means the specific engineered features should not be treated as universally exchangeable. Quest difficulty, post-failure persistence, and the other retained features are best understood as context-specific operationalizations of broader behavioral ideas. For that reason, the main contribution that should generalize from this study is the design process for identifying and engineering useful proxy features, not the exact feature set itself. A useful next step would be a model showing which types of game mechanics or design elements map most clearly onto specific constructs.

Another limitation is the level of aggregation used in the modeling. The current models rely on participant-level summary features, which compress within-session variation into a single profile for each player. That reduction was useful for a first-pass predictive study, but it also removes

much of the moment-to-moment structure that may matter most for a construct like self-efficacy. The weaker posttest target should be interpreted in that light. It is not only a lower-performing target; it is also evidence that a static summary has a ceiling once the construct itself begins to move during the experience. In other words, when self-efficacy shifts during play, a participant-level aggregate may no longer align cleanly with the belief state measured at the end. Thus, future modeling should prioritize models that work with long-form data (e.g., Hidden Markov Models, Dynamic Bayesian Networks).

Finally, the gameplay environment itself was limited. *Mean Alchemy* functioned as a minimum viable product within a short study window, and that likely constrained both the richness and the duration of the behavioral traces available for modeling. Some constructs, especially risk-taking, may require a setting with more meaningful stakes, more repeated opportunities to act, or a longer period of engagement before their behavioral signatures become clear. Longer-term deployments or richer existing game environments would likely produce stronger traces, more stable feature estimates, and a better test of how well self-efficacy can be inferred from behavior over time.

6.3.2 Implications

The implications for measurement and assessment design are clear. My study suggests that there is enough evidence in behaviors associated with self-efficacy to begin building models that can infer those beliefs without relying exclusively on self-report. There is ample evidence to support concerns about self-report bias in self-efficacy measures (Bandura, 1997; Choi et al., 2023; Kruger & Dunning, 1999). The results of my study suggest that it is possible to detect a signal of self-efficacy beliefs from behavior, which means that the field should continue moving toward more objective measures of these internal *black box* beliefs. That is an important step forward because it opens the door to more accurate and equitable assessments of self-efficacy—and other hard-to-measure constructs—which can inform instruction and support for students. By reducing reliance on self-report, or even on explicit assessments such as quizzes and surveys, future systems may be able to reduce burden on students and teachers while also providing more timely and actionable data.

In terms of modeling, the results in Figure 5.10 support the importance of considering the dynamic nature of self-efficacy beliefs. Future work should prioritize models that can capture temporal patterns in behavior, especially around moments of success and failure, rather than relying solely on participant-level aggregates. Analyses that were previously impractical are becoming more feasible with modern computational tools and larger datasets. Thus, the field should move toward

models that can capture the dynamic interplay between behavior and beliefs over time, which is likely to provide a richer and more accurate understanding of constructs like self-efficacy—especially in game-based contexts, where design can elicit behaviors that are closely tied to those beliefs.

However, teachers and practitioners may receive the greatest benefit from this unobtrusive and continuous measurement of self-efficacy beliefs and related self-beliefs. If a system can detect when a student's self-efficacy is low, it could provide timely support or interventions to help strengthen those beliefs and improve learning outcomes. For example, if a student is struggling with a particular concept and their self-efficacy is dropping, the system could offer additional resources, encouragement, or even adjust the difficulty of the material to help them succeed. This kind of responsive support could be especially valuable in online or blended learning environments where teachers may not have as much direct contact with students. By reducing interruptions to instruction and practice for quizzes and tests, teachers could focus more on supporting students and spend less time administering, grading, and reporting on assessments. That is not to say that all assessments should be replaced with behavioral inference models, but rather that there is an opportunity to use these models to complement traditional assessments and provide a more holistic picture of students' learning and development.

Looking ahead, future research should test whether this same design logic can be extended into broader learning environments and more adaptive forms of assessment. Generative artificial intelligence may eventually make it more practical to build learning experiences that are responsive to content, classroom context, and individual students while also producing richer behavioral data. If that is the direction the field takes, then the central challenge will not only be technical. It will be ensuring that those systems are designed to support instruction, reduce burden on teachers, and use student data ethically, safely, and effectively. This dissertation is intended as one contribution to that next stage of work.

APPENDIX A

IRB DOCUMENTS

FSU | DIVISION OF RESEARCH

Human Subjects Protection · Institutional Review Board

APPROVAL

May 21, 2025

Gerald Fulwider

[redacted]

Dear Gerald Fulwider:

On 5/21/2025, the IRB reviewed the following submission:

Type of Review:	Expedited (5) Data, documents, records, or specimens; (6) Voice, video, digital, or image recordings; (7)(a) Behavioral research; (8)(c) Data analysis
Title:	Assessing Self-Efficacy Through Play: A Game-Based Approach to Measuring the Unobservable
Investigator:	Gerald Fulwider
Submission ID:	STUDY00006193
Study ID:	STUDY00006193
Funding:	None
IND, IDE, or HDE:	None
Documents Reviewed:	• Gameplay Opt-out Form, Category: Consent Form; • Gameplay Screenshots, Category: Other; • HRP-503a, Category: IRB Protocol; • Interview Invitation, Category: Recruitment Materials; • Self-report Scales, Category: Survey/Questionnaire; • Study Information Sheet, Category: Information Sheet;

The IRB approved the protocol, effective from 5/21/2025.

Other Information: You are advised that any change or revision to the protocol for this project must, through a study modification, be reviewed and approved by the IRB prior to implementation of the proposed modification(s).

Federal regulations require that the Principal Investigator promptly report, through a Report of New Information, any incident involving, for example, the following: a new or increased risk or safety issue; harm experienced by a study participant; non-compliance with federal regulations or the determinations of the IRB; audits, monitoring reports or inspections by study sponsors, monitors or federal agencies; breach of confidentiality; complaint of a study participant; etc.) (see the Investigator Manual (HRP-103), which can be found in RAMP IRB, under the IRB, Library and General tabs).

INFORMATION ABOUT THE FSU GAME-BASED LEARNING RESEARCH STUDY
UNDERSTANDING STUDENT CONFIDENCE THROUGH GAMEPLAY

I am a doctoral candidate at Florida State University studying how students build confidence in their academic abilities—a concept called **self-efficacy**. Research shows that students who believe in their ability to succeed are more likely to stay engaged, try harder, and perform better in school. But traditional measures—like surveys—don't always capture how students actually respond to challenges in real time.

My study explores whether a game-based approach can offer a better way to measure self-efficacy, by observing how students behave while solving math problems in a story-driven educational video game. This research is part of a doctoral dissertation approved by the FSU Institutional Review Board.

Surveys: Students who participate in the study will complete a brief set of digital surveys before and after playing the game. These surveys include:

- A self-efficacy questionnaire about how confident they feel solving math problems.
- A short knowledge quiz related to the game's content.
- A demographic form that asks basic background questions (e.g., grade level, prior math experience).

Gameplay: After completing the surveys, students will play an educational video game called *Mean Alchemy*, designed specifically for middle schoolers. The game follows a young apprentice as they learn to become a data "alchemist," using data to empower their magical companion. As students play, they will practice key statistical concepts like mean, standard deviation, and creating histograms all within an age-appropriate narrative.

Interviews: Following gameplay, I will invite students to participate in a short follow-up interview. These interviews are completely voluntary and will be offered to a small number of students whose survey responses and gameplay data show patterns of interest. Interviews will last about 20–30 minutes and will be conducted via Zoom. Families can choose a time that works best for them. Students who complete the interview will receive a $10 digital gift card (e.g., Amazon or Target) as a thank-you for their time.

Intervention in the Classroom: Surveys and gameplay will be completed during regular school hours over the course of 2–3 class periods. Teachers will help identify convenient times for participation. If a student is invited to participate in a follow-up interview, that portion will take place outside of school hours.

If ou have an uestions or would like more informatio [redacted] lwider [redacted] r my advisor, Dr. Bret Staudt Willet [redacted]

The Florida State University Institutional Review Board ("IRB") is overseeing this research. The FSU IRB is a group of people who perform official independent review of research studies before studies begin to ensure that the rights and welfare of participants are protected. If you have questions about your rights or wish to speak with someone other than the research team, you may contact the Florida State University IRB:

[redacted]

PARENTAL RESEARCH OPT-OUT FORM

Title of Research Study: Game-based assessment of statistical self-efficacy: An alternative to the self-report of non-cognitive beliefs

Principal Investigator: G.Curt Fulwider
Doctoral Candidate of Instructional Systems at FSU
[redacted]

Student Advisor: Dr. Bret Staudt Willet
Assistant Prof. of Instructional Systems at FSU
[redacted]

Your child is being asked to participate in this study because they are in one of the 6th to 8th grade classes available. This study is asking students to complete a statistical self-efficacy and content knowledge survey before and after playing an age-appropriate digital learning game. They may also be invited to participate in interviews outside of class time if their results show patterns of interest. If at any time your child wishes to no longer participate, they will be allowed to quit and have their data destroyed without prejudice.

The goal of my study is to explore whether gameplay-related behavioral patterns may be used to assess their statistical self-beliefs. We will retain the survey data, gameplay data, and interview transcripts (if they participate in interviews). The only tie to identifying information will be a unique, randomly assigned identification number. The key to their protected information (i.e., name and student number) will be available only to the principal investigator and destroyed once interviews are completed. The identifying information is only used to extend interview invitations.

This study has been approved by the FSU Human Subjects Committee at Florida State University and it has been determined that the study poses minimal risk to students—no more risk than students would encounter on a typical day at school. Thus, allowing us to provide the parental opt-out screener.

We only need this form returned from you if you do NOT want your child to participate in the study. Interview invitations will be sent out post gameplay and you may decline the interviews at a late date. **If you are ok with your child completing the surveys and playing the game in class you do not need to return this form.**

If you have any questions or complaints about the study, you can talk to principal investigator, Curt Fulwider, at the [redacted] approved this study.

I have read and considered the information presented in this form. My signature and the date below indicate that I DO NOT *want my child to* complete the surveys nor participate in gameplay.

Child's name (Print Clearly):

APPENDIX B

SELF-EFFICACY INSTRUMENT

The self-efficacy instrument was administered at both pretest and posttest. For each item, students used a slider from 0 to 100, where 0 indicated *I cannot do this at all* and 100 indicated *I am completely certain I can do this*. Students were instructed to move the slider anywhere between 0 and 100 to best match their confidence level.

Note. This instrument was adapted from the statistical literacy self-efficacy scale reported by Carmichael and Hay (2009).

1. Solve problems that use averages.
2. Find when a newspaper article has used the wrong type of average.
3. Explain to a friend how probability (or chance) is calculated.
4. Show data correctly on a histogram.
5. Explain the meaning of a graph in a newspaper or on the internet.
6. Find a mistake in someone else's graph.
7. Explain when conclusions that are based on surveys might be wrong.
8. Arrange my data correctly into a table.
9. Explain how to select a fair sample of students for a school survey.

APPENDIX C

CONTENT PRETEST INSTRUMENT

The content pretest was administered before gameplay. The item wording below reflects the prompts stored in the Qualtrics survey export used for the study.

1. The numbers 2, 4, 6, 8, and 10 are shown on a number line. Without doing the exact math, what would be a good estimate for the mean of this group?
2. Which of these sets of numbers is most spread out?
3. Which measure tells you the middle value when the data is in order?
4. Two sets of numbers have the same mean. Which would have a higher standard deviation?
5. Which of the following sets has a mean of 10 and the lowest standard deviation?
6. A data set has a low standard deviation. What does that mean?
7. Which of the following would increase the mean the most?
8. If the mean is 50 and the standard deviation is 5, which number is farthest from the mean?
9. Choose the histogram that best represents a dataset with: Mean = 5, SD = 0.5.
10. Look at this histogram. Estimate the mean and standard deviation.

APPENDIX D

CONTENT POSTTEST INSTRUMENT

The content posttest was administered after gameplay as an alternate form of the content measure. The item wording below reflects the prompts stored in the Qualtrics survey export used for the study.

1. The numbers 3, 5, 7, 9, and 11 are shown on a number line. Without doing the exact math, what would be a good estimate for the mean?
2. Which of these data sets shows the most variation?
3. Which measure tells you how far numbers are from the average?
4. Two data sets have the same median. Which one will have the higher standard deviation?
5. Which of the following sets has a mean of 5 and the lowest standard deviation?
6. A data set has a high standard deviation. What does that mean?
7. Which of the following would decrease the mean the most?
8. If the mean is 30 and the standard deviation is 4, which number is farthest from the mean?
9. Choose the histogram that best represents a dataset with: Mean = 7, SD = 2.
10. Look at this histogram. Estimate the mean and standard deviation.

APPENDIX E

MOTIVATION SURVEY INSTRUMENT

The motivation survey used a 7-point agreement scale for all items, where 1 indicated *Strongly Disagree* and 7 indicated *Strongly Agree*.

E.1 Interest in Math and Statistics

- Math is a fun subject to study.
- I enjoy learning about statistics.
- I think math is boring.*
- I think statistics is pointless.*

E.2 Interest in Video Games

- I consider myself to be "a gamer".
- I enjoy playing video games in my free time.
- I don't like playing video games.*
- Playing video games feels like a waste of time.*

E.3 Motivation for Study Participation

- I am excited to participate in this study.
- I am only participating in this study because I have to.*
- I am participating in this study because I think it will be interesting.
- I would be doing something else right now.*

Note. Items marked with an asterisk were reverse coded before scale construction.

REFERENCES

Almond, R., Shute, V. J., Tingir, S., & Rahimi, S. (2020). Identifying observable outcomes in game-based assessments. In *Innovative psychometric modeling and methods* (pp. 163–192). Information Age Publishing, Inc.

Appelbaum, S. H., & Hare, A. (1996). Self-efficacy as a mediator of goal setting and performance: Some human resource applications. *Journal of Managerial Psychology*, *11*(3), 33–47. https://doi.org/10.1108/02683949610113584

Baker, R. S. (2019). Challenges for the future of educational data mining: The baker learning analytics prizes [Publisher: Zenodo Version Number: 1.0.0]. https://doi.org/10.5281/ZENODO.3554745

Baker, R. S., & Inventado, P. S. (2014). Educational Data Mining and Learning Analytics. In J. A. Larusson & B. White (Eds.), *Learning Analytics* (pp. 61–75). Springer New York. https://doi.org/10.1007/978-1-4614-3305-7_4

Baker, R. S., Walonoski, J. A., Heffernan, N. T., Roll, I., Corbett, A., & Koedinger, K. R. (2008). Why students engage in "gaming the system" behavior in interactive learning environments. *Journal of Interactive Learning Research*, *19*(2), 185–224. https://pact.cs.cmu.edu/koedinger/pubs/Baker,%20R.,%20Walonoski,%20J.A.,%20Heffernan,%20N.T.,%20Roll,%20I.%20Corbett,%20A.,%20Koedinger,%20K.R..pdf

Bandura, A. (1977). Self-efficacy: Toward a unifying theory of behavioral change. *Psychological Review*, *84*(2), 191–215. https://doi.org/10.1037/0033-295X.84.2.191

Bandura, A. (1986). *Social foundations of thought and action: A social cognitive theory*. Prentice-Hall.

Bandura, A. (Ed.). (1995). *Self-efficacy in changing societies*. Cambridge University Press.

Bandura, A. (1997). *Self-efficacy: The exercise of control*. W. H. Freeman.

Bandura, A. (2006). Toward a psychology of human agency. *Perspectives on Psychological Science*, *1*(2), 164–180. https://doi.org/10.1111/j.1745-6916.2006.00011.x

Bandura, A. (2012). On the functional properties of perceived self-efficacy revisited [Place: Los Angeles, CA Publisher: SAGE Publications]. *Journal of management*, *38*(1), 9–44. https://doi.org/10.1177/0149206311410606

Bandura, A., & Adams, N. E. (1977). Analysis of self-efficacy theory of behavioral change. *Cognitive Therapy and Research*, *1*(4), 287–310. https://doi.org/10.1007/BF01663995

Bandura, A., & Locke, E. A. (2003). Negative self-efficacy and goal effects revisited. *Journal of Applied Psychology*, *88*(1), 87–99. https://doi.org/10.1037/0021-9010.88.1.87

Banilower, E. R., Smith, P. S., Weiss, I. R., Malzahn, K. A., Campbell, K. M., & Weis, A. M. (2013). *Report of the 2012 national survey of science and mathematics education*. Horizon Research Inc. Chapel Hill, NC.

Beghetto, R. A. (2009). Correlates of intellectual risk taking in elementary school science. *Journal of Research in Science Teaching*, *46*(2), 210–223. https://doi.org/10.1002/tea.20270

Ben-Nun, P. (2008). Respondent fatigue. In *Encyclopedia of survey research methods* (p. 743). Sage Publications, Inc. Retrieved September 17, 2024, from https://methods.sagepub.com/reference/encyclopedia-of-survey-research-methods/n480.xml

Black, P., & Wiliam, D. (2010). Inside the black box: Raising standards through classroom assessment. *Phi Delta Kappan*, *92*(1), 81–90. https://doi.org/10.1177/003172171009200119

Bouffard, T., Bouchard, M., Goulet, G., Denoncourt, I., & Couture, N. (2005). Influence of achievement goals and self-efficacy on students' self-regulation and performance. *International Journal of Psychology*, *40*(6), 373–384. https://doi.org/10.1080/00207590444000302

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Carmichael, C., & Hay, I. (2009). The development and validation of the students' self-efficacy for statistical literacy scale. *MERGA 32: Crossing divides*, *1*, 97–104.

Cheuk, T. (2021). Can AI be racist? color-evasiveness in the application of machine learning to science assessments. *Science Education*, *105*(5), 825–836. https://doi.org/10.1002/sce.21671

Choi, H., Winne, P. H., & Brooks, C. (2023). Reconfiguring Measures of Motivational Constructs Using State-Revealing Trace Data. In V. Kovanovic, R. Azevedo, D. C. Gibson, & D. Lfenthaler (Eds.), *Unobtrusive Observations of Learning in Digital Environments* (pp. 73–89). Springer International Publishing. https://doi.org/10.1007/978-3-031-30992-2_5

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, *20*(1), 37–46. https://doi.org/10.1177/001316446002000104

Congdon, E., Bato, A. A., Schonberg, T., Mumford, J. A., Karlsgodt, K. H., Sabb, F. W., London, E. D., Cannon, T. D., Bilder, R. M., & Poldrack, R. A. (2013). Differences in neural activation as a function of risk-taking task parameters. *Frontiers in Neuroscience*, *7*. https://doi.org/10.3389/fnins.2013.00173

Czerwiński, S. K., Konarski, R., & Atroszko, P. A. (2025). Lack of measurement invariance in mental health assessment across intelligence levels: Investigation into nonlinearity reveals a broader issue. *Intelligence*, *113*, 1–13. https://doi.org/10.1016/j.intell.2025.101963

Deitzer, J. R., Leban, L., Copes, H., & Wilcox, S. (2021). Criminal self-efficacy and perceptions of risk and reward among women methamphetamine manufacturers [MAG ID: 3150073157]. *Justice Quarterly*, 1–24. https://doi.org/10.1080/07418825.2021.1901965

Delandshere, G. (2002). Assessment as inquiry. *Teachers College Record: The Voice of Scholarship in Education*, *104*(7), 1461–1484. https://doi.org/10.1111/1467-9620.00210

Denovan, A., Dagnall, N., & Drinkwater, K. (2023). Examining what mental toughness, ego resiliency, self-efficacy, and grit measure: An exploratory structural equation modelling bifactor approach. *Current Psychology*, *42*(26), 22148–22163. https://doi.org/10.1007/s12144-022-03314-5

DiBenedetto, M. K., & Schunk, D. H. (2022). Assessing Academic Self-efficacy. In M. S. Khine & T. Nielsen (Eds.), *Academic Self-efficacy in Education* (pp. 11–37). Springer Singapore. https://doi.org/10.1007/978-981-16-8240-7_2

DiCerbo, K. E. (2014). Game-based assessment of persistence. *Journal of Educational Technology & Society*, *17*(1), 17–28. https://doi.org/www.jstor.org/stable/jeductechsoci.17.1.17

Eisenberger, R. (1992). Learned industriousness. *Psychological Review*, *99*(2), 248–267. https://doi.org/10.1037/0033-295X.99.2.248

Feather, N. T. (1962). The study of persistence. *Psychological Bulletin*, *59*(2), 94–115. https://doi.org/10.1037/h0042645

Fisher, A., Rudin, C., & Dominici, F. (2019). All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously. *Journal of Machine Learning Research*, *20*(177), 1–81. https://jmlr.org/papers/v20/18-760.html

Fraine, N., & McDade, R. (2009). Reducing bias in psychometric assessment of culturally and linguistically diverse students from refugee backgrounds in australian schools: A process approach. *Australian Psychologist*, *44*(1), 16–26. https://doi.org/10.1080/00050060802582026

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, *29*(5), 1189–1232. https://doi.org/10.1214/aos/1013203451

Gonzalez-DeHass, A., Furner, J., Vasquez-Colina, M., & Morris, J. (2022). Achievement goals as predictors of female pre-service elementary school teachers' self-efficacy for learning math in a methods course. *Journal of Early Childhood Teacher Education*, *43*(4), 568–587. https://doi.org/10.1080/10901027.2021.1955052

Guskey, T. R., & Link, L. J. (2019). Exploring the factors teachers consider in determining students' grades. *Assessment in Education: Principles, Policy & Practice*, *26*(3), 303–320. https://doi.org/10.1080/0969594X.2018.1555515

Gutman, L., & Schoon, I. (2013, November 21). *The impact of non-cognitive skills on outcomes for young people* (Literature Review). The Institute of Education.

Hand, D. J., & Yu, K. (2001). Idiot's bayes—not so stupid after all? *International Statistical Review*, *69*(3), 385–398.

Huang, C. (2016). Achievement goals and self-efficacy: A meta-analysis. *Educational Research Review*, *19*, 119–137. https://doi.org/10.1016/j.edurev.2016.07.002

Hutchinson, B., & Mitchell, M. (2019). 50 years of test (un)fairness: Lessons for machine learning. *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 49–58. https://doi.org/10.1145/3287560.3287600

Kates, S., Paulsen, T., Yntiso, S., & Tucker, J. A. (2022). Bridging the grade gap: Reducing assessment bias in a multi-grader class. *Political Analysis*, 1–9. https://doi.org/10.1017/pan.2022.27

Kormos, C., & Gifford, R. (2014). The validity of self-report measures of proenvironmental behavior: A meta-analytic review. *Journal of Environmental Psychology*, *40*, 359–371. https://doi.org/10.1016/j.jenvp.2014.09.003

Kostick-Quenet, K. M., Cohen, I. G., Gerke, S., Lo, B., Antaki, J., Movahedi, F., Njah, H., Schoen, L., Estep, J. E., & Blumenthal-Barby, J. (2022). Mitigating racial bias in machine learning. *Journal of Law, Medicine & Ethics*, *50*(1), 92–100. https://doi.org/10.1017/jme.2022.13

Krueger, N., & Dickson, P. R. (1994). How believing in ourselves increases risk taking: Perceived self-efficacy and opportunity recognition. *Decision Sciences*, *25*(3), 385–400. https://doi.org/10.1111/j.1540-5915.1994.tb01849.x

Kruger, J., & Dunning, D. (1999). Unskilled and unaware of it: How difficulties in recognizing one's own incompetence lead to inflated self-assessments. *Journal of Personality and Social Psychology*, *77*(6), 1121–1134. https://doi.org/10.1037/0022-3514.77.6.1121

Kyllonen, P. C. (2020). Advances in measuring "hard-to-measure" skills. In M. S. Khine (Ed.), *Contemporary perspectives on research in educational assessment*. Information Age Publishing, Inc.
OCLC: 1141255662.

Kyllonen, P. C., Walters, A. M., & Kaufman, J. C. (2005). Noncognitive constructs and their assessment in graduate education: A review. *Educational Assessment*, *10*(3), 153–184. https://doi.org/10.1207/s15326977ea1003_2

Lee, J.-E., & Recker, M. (2017). *Measuring Students' Use of Self-Regulated Learning Strategies from Learning Management System Data: An Evidence-Centered Design Approach About Analytics for Learning (A4L)* (Technical Report). Utah State University. https://doi.org/10.13140/RG.2.2.24971.75047

Lent, R. W., Brown, S. D., & Larkin, K. C. (1984). Relation of self-efficacy expectations to academic achievement and persistence. *Journal of Counseling Psychology*, *31*(3), 356–362. https://doi.org/10.1037/0022-0167.31.3.356

Liao, H.-A., Edlin, M., & Ferdenzi, A. C. (2014). Persistence at an urban community college: The implications of self-efficacy and motivation. *Community College Journal of Research and Practice*, *38*(7), 595–611. https://doi.org/10.1080/10668926.2012.676499

Lira, B., O'Brien, J. M., Peña, P. A., Galla, B. M., D'Mello, S., Yeager, D. S., Defnet, A., Kautz, T., Munkacsy, K., & Duckworth, A. L. (2022). Large studies reveal how reference bias limits policy applications of self-report measures. *Scientific Reports*, *12*(1), 19189. https://doi.org/10.1038/s41598-022-23373-9

Liu, T., & Israel, M. (2022). Uncovering students' problem-solving processes in game-based learning environments. *Computers & Education*, *182*, 104462. https://doi.org/10.1016/j.compedu.2022.104462

Locke, E. A., Frederick, E., Lee, C., & Bobko, P. (1984). Effect of self-efficacy, goals, and task strategies on task performance. *Journal of Applied Psychology*, *69*(2), 241–251. https://doi.org/https://psycnet.apa.org/doi/10.1037/0021-9010.69.2.241

Locke, E. A., & Latham, G. P. (1991). A theory of goal setting & task performance. *The Academy of Management Review*, *16*.

Locke, E. A., & Latham, G. P. (2002). Building a practically useful theory of goal setting and task motivation: A 35-year odyssey. *American Psychologist*, *57*(9), 705–717. https://doi.org/10.1037/0003-066X.57.9.705

Long, P., & Siemens, G. (2011). Penetrating the fog: Analytics in learning and education. *EDUCAUSE Review*, *46*(5), 31–40. https://er.educause.edu/articles/2011/9/penetrating-the-fog-analytics-in-learning-and-education

McIntyre, M. M. (2023). Statistics is puzzling: Testing a novel approach to statistics learning [Publisher: Educational Publishing Foundation]. *Scholarship of Teaching and Learning in Psychology*, *9*(2), 150–158. https://doi.org/10.1037/stl0000204

McQuiggan, S. W., Mott, B. W., & Lester, J. C. (2008). Modeling self-efficacy in intelligent tutoring systems: An inductive approach. *User Modeling and User-Adapted Interaction*, *18*(1–2), 81–123. https://doi.org/10.1007/s11257-007-9040-y

Merritt, C. J., & Tharp, I. J. (2013). Personality, self-efficacy and risk-taking in parkour (free-running) [MAG ID: 1970840496]. *Psychology of Sport and Exercise*, *14*(5), 608–611. https://doi.org/10.1016/j.psychsport.2013.03.001

Mislevy, R. J. (1993). Foundations of a new test theory. In N. Frederiksen, R. J. Mislevy, & I. I. Bejar (Eds.), *Test theory for a new generation of tests.* (pp. 19–35). Lawrence Erlbaum Associates, Inc.

Mislevy, R. J. (2003). Substance and structure in assessment arguments. *Law, Probability and Risk*, *2*(4), 237–258. https://doi.org/https://Doi.org/10.1093/Lpr/2.4.237

Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2003). Focus article: On the structure of educational assessments. *Measurement: Interdisciplinary Research & Perspective*, *1*(1), 3–62. https://doi.org/10.1207/S15366359MEA0101_02

Montford, W., & Goldsmith, R. E. (2016). How gender and financial self-efficacy influence investment risk taking: Investment risk taking. *International Journal of Consumer Studies*, *40*(1), 101–106. https://doi.org/10.1111/ijcs.12219

Pellegrino, J. W., Baxter, G. P., & Glaser, R. (1999). Addressing the "two disciplines" problem: Linking theories of cognition and learning with assessment and instructional practice. *Review of Research in Education*, *24*(1), 307–353. https://doi.org/10.3102/0091732X024001307

Powers, D. M. W. (2020). Evaluation: From precision, recall and f-measure to ROC, informedness, markedness and correlation [Version 1]. *arXiv*. https://doi.org/10.48550/ARXIV.2010.16061

Roick, J., & Ringeisen, T. (2017). Self-efficacy, test anxiety, and academic success: A longitudinal validation. *International Journal of Educational Research*, *83*, 84–93. https://doi.org/10.1016/j.ijer.2016.12.006

Rosenman, R., Tennekoon, V., & Hill, L. G. (2011). Measuring bias in self-reported data. *International Journal of Behavioural and Healthcare Research*, *2*(4), 320. https://doi.org/10.1504/IJBHR.2011.043414

Rosenthal, D., Moore, S., & Flynn, I. (1991). Adolescent self-efficacy, self-esteem and sexual risk-taking. *Journal of Community & Applied Social Psychology*, *1*(2), 77–88. https://doi.org/10.1002/casp.2450010203

Rowe, E., Almeda, M. V., Asbell-Clarke, J., Scruggs, R., Baker, R. S., Bardar, E., & Gasca, S. (2021). Assessing implicit computational thinking in zoombinis puzzle gameplay. *Computers in Human Behavior*, *120*, 106707. https://doi.org/10.1016/j.chb.2021.106707

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, *1*(5), 206–215. https://doi.org/10.1038/s42256-019-0048-x

Salinger, J. D. (1951). *The catcher in the rye*. Little, Brown; Company.

Schunk, D. H. (1984). Self-efficacy perspective on achievement behavior. *Educational Psychologist*, *19*(1), 48–58. https://doi.org/10.1080/00461528409529281

Schunk, D. H., & Pajares, F. (2002). The development of academic self-efficacy. In A. Wigfield & J. S. Eccles (Eds.), *Development of achievement motivation* (pp. 15–31). Academic Press. https://doi.org/10.1016/B978-012750053-9/50003-6

Shute, V. J. (2008). Focus on formative feedback. *Review of Educational Research*, *78*(1), 153–189. https://doi.org/10.3102/0034654307313795

Shute, V. J. (2011). Stealth assessment in computer-based games to support learning. In S. Tobias & J. Fletcher (Eds.), *Computer games and instruction* (pp. 503–524). Information Age Publishing, Inc.

Shute, V. J., & Becker, B. J. (2010). Prelude: Assessment for the 21st century. In V. J. Shute & B. J. Becker (Eds.), *Innovative assessment for the 21st century* (pp. 1–11). Springer US. https://doi.org/10.1007/978-1-4419-6530-1_1

Shute, V. J., Masduki, I., & Donmez, O. (2010). Conceptual framework for modeling, assessing and supporting competencies within game environments. *Cognition and Learning*, *8*, 137–161.

Shute, V. J., & Rahimi, S. (2021). Stealth assessment of creativity in a physics video game. *Computers in Human Behavior*, *116*, 106647. https://doi.org/10.1016/j.chb.2020.106647

Shute, V. J., Rahimi, S., Smith, G., Ke, F., Almond, R., Dai, C.-P., Kuba, R., Liu, Z., Yang, X., & Sun, C. (2021). Maximizing learning without sacrificing the fun: Stealth assessment, adaptivity and learning supports in educational games. *Journal of Computer Assisted Learning*, *37*(1), 127–141. https://doi.org/10.1111/jcal.12473

Skinner, B. F. (1976). *About behaviorism*. Vintage Books.

Slater, S., Bowers, A., Kai, S., & Shute, V. J. (2017). A typology of players in the game physics playground. *Proceedings of DiGRA 2017 Conference*. https://doi.org/10.26503/dl.v2017i1.925

Stankov, L., & Lee, J. (2014). Juxtaposing math self-efficacy and self-concept as predictors of long-term achievement outcomes. *Educational Psychology*, *34*(1), 29–48. https://doi.org/10.1080/01443410.2013.797339

Stankov, L., Morony, S., & Lee, Y. P. (2014). Confidence: The best non-cognitive predictor of academic achievement? *Educational Psychology*, *34*(1), 9–28. https://doi.org/10.1080/01443410.2013.814194

Steegen, S., Tuerlinckx, F., Gelman, A., & Vanpaemel, W. (2016). Increasing transparency through a multiverse analysis. *Perspectives on Psychological Science*, *11*(5), 702–712. https://doi.org/10.1177/1745691616658637

Syal, S., & Nietfeld, J. L. (2020). The impact of trace data and motivational self-reports in a game-based learning environment. *Computers & Education*, *157*, 103978. https://doi.org/10.1016/j.compedu.2020.103978

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, *58*(1), 267–288. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x

Ventura, M., & Shute, V. J. (2013). The validity of a game-based assessment of persistence. *Computers in Human Behavior*, *29*(6), 2568–2572. https://doi.org/10.1016/j.chb.2013.06.033

Ventura, M., Shute, V. J., & Small, M. (n.d.). CHAPTER 8 – assessing persistence in educational games. *2*.

von der Embse, N., Jester, D., Roy, D., & Post, J. (2018). Test anxiety effects, predictors, and correlates: A 30-year meta-analytic review. *Journal of Affective Disorders*, *227*, 483–493. https://doi.org/10.1016/j.jad.2017.11.048

Wang, J., Stebbins, A., & Ferdig, R. E. (2022). Examining the effects of students' self-efficacy and prior knowledge on learning and visual behavior in a physics game. *Computers & Education*, *178*, 104405. https://doi.org/10.1016/j.compedu.2021.104405

Watson, K., Baranowski, T., Thompson, D., Jago, R., Baranowski, J., & Klesges, L. M. (2006). Innovative application of a multidimensional item response model in assessing the influence of social desirability on the pseudo-relationship between self-efficacy and behavior. *Health Education Research*, *21*, i85–i97. https://doi.org/10.1093/her/cyl137

Wood, R., & Bandura, A. (1989). Social cognitive theory of organizational management. *The Academy of Management Review*, *14*(3), 361. https://doi.org/10.2307/258173

Wu, F., Fan, W., Arbona, C., & De La Rosa-Pohl, D. (2020). Self-efficacy and subjective task values in relation to choice, effort, persistence, and continuation in engineering: An expectancy-value theory perspective. *European Journal of Engineering Education*, *45*(1), 151–163. https://doi.org/10.1080/03043797.2019.1659231

Zhou, M., & Winne, P. H. (2012). Modeling academic achievement by self-reported versus traced goal orientation. *Learning and Instruction*, *22*(6), 413–419. https://doi.org/10.1016/j.learninstruc.2012.03.004

Zimmerman, B. J. (2000). Self-efficacy: An essential motive to learn. *Contemporary Educational Psychology*, *25*(1), 82–91. https://doi.org/10.1006/ceps.1999.1016

Zimmerman, B. J., Bandura, A., & Martinez-Pons, M. (1992). Self-motivation for academic attainment: The role of self-efficacy beliefs and personal goal setting. *American Educational Research Journal*, *29*(3), 663–676. https://doi.org/10.3102/00028312029003663

BIOGRAPHICAL SKETCH

G. Curt Fulwider grew up in a small town in western Kansas and completed his undergraduate studies at Washburn University in Topeka. After graduation, he accepted a position teaching English at a university in southern China. His time there broadened his academic interests and eventually led him to Tallahassee, where he pursued graduate study at Florida State University. During his graduate work at Florida State University, Fulwider completed two master's degrees before entering doctoral candidacy. The second emerged from a growing interest in statistics, research design, and educational measurement, which became central to his academic work. His research focuses on game-based assessment, stealth assessment, self-efficacy, educational data mining, learning analytics, and the use of machine learning to support measurement in education. He plans to continue this work in future academic and research settings.