THE EFFECT OF PERCEIVED RELEVANCE OF DIGITAL BADGES ON STUDENT ENGAGEMENT

by

Ross Higashi

BS, Logic and Computation, Carnegie Mellon University, 2007

Submitted to the Graduate Faculty of

the School of Education in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

University of Pittsburgh

2018

UNIVERSITY OF PITTSBURGH

SCHOOL OF EDUCATION

This dissertation was presented

by

Ross Higashi

It was defended on

July 20, 2018

and approved by

Dr. Thomas Akiva, Assistant Professor, Learning Sciences and Policy

Dr. Lindsay Page, Assistant Professor, Learning Sciences and Policy

Dr. Amy Ogan, Assistant Professor, Human-Computer Interaction Institute, Carnegie Mellon University

Dissertation Advisor: Dr. Christian Schunn, Professor, Learning Sciences and Policy

THE EFFECT OF PERCEIVED RELEVANCE OF DIGITAL BADGES ON STUDENT ENGAGEMENT

Ross Higashi, PhD

University of Pittsburgh, 2018

Open digital badge systems have been promoted as potentially impactful interventions in education, but past studies have found learning and motivational effects that vary drastically by learner, and it is unclear what elements matter. This may be due to a failure to account for learners' subjective evaluations of the badges themselves, which likely moderate the badges' impact. I propose a theoretical model which unpacks the traditional "black box" view, shifting the focus from *effects of digital badges* to *processes by which badges affect learners*. In the initial model, learners' *subjective evaluation of badges* determines their *engagement with badges*, which in turn influences their *engagement with program activities*. If there is such an effect, I am interested in whether it is a general mechanism across contexts; whether its effects are equitable with respect to age, sex, and race; and whether it may inform the design of better badging systems in the future.

Chapter 1 situates this work in relation to previous digital badge research, and introduces a synthetic badge-facing factor called Perceived Badge Relevance (PBR) that represents the degree to which a learner finds a program's badges "relevant" to them: whether the learner thinks the badges make sense, thinks they are valuable, and wants to earn them. Chapter 2 uses

PBR to test for *subjective evaluation of badges* effects on learner engagement across 45 summer programs, finding an overall positive relationship between PBR and engagement that does not vary between programs. Shifting to a design-based research mode, Chapter 3 describes the design for a robust badge evidence system, used by the badge system in Chapter 4, which tests for a longitudinal effect of PBR in an online course. This study finds that PBR predicts rank-order shifts in engagement over time, is not “explained away” by long-term interest or demographics, and is neutral with respect to age, sex, and minoritized racial status. Engagement also predicts relative change in PBR, indicating the potential for a positive feedback loop. Chapter 5 synthesizes conclusions across studies, evaluates support for the *subjective evaluation of badges* hypothesis, and puts forth a continuing research agenda.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ACKNOWLEDGMENTS

Dedicated to my colleagues in the LSAP program, its endlessly supportive faculty, and everyone who makes LRDC the wonderful place it is. Thank you of course to my advisor, Chris Schunn, and to my committee. This work would not have been possible without your guidance. Special thanks to the folks at the Carnegie Mellon Robotics Academy for their hard work on the technical side. And finally, to my future-famous friends in 823: never stop doing you.

1.0 CHAPTER 1: INTRODUCTION

A badge is a marker of accomplishment, descended from marks of pilgrimage and political affiliation in the Middle Ages, military medals, and featuring prominently in modern-day Scouting movements (Ostashewski & Reid, 2015). Badges in digital format are often associated with video game achievements, but in the last few years have been re-envisioned as a potential avenue by which real-world skills might be communicated to learners, potential employers, and many other players in the world of education and credentialing. “Open” badge systems in particular – sociotechnical ecosystems without burdensome restrictions on who can issue badges – have been proposed as a catalyst for a wholesale change in the world of education (MacArthur Foundation, 2011; Mozilla, 2015).

One major area in which badges have been presumed to have use is in motivating students to engage in learning activities. Studies have investigated whether, when, and for whom badges work. However, there is a gap in the literature (Chapter 1) around *how* badges work. Specifically, there is not a strong theoretical accounting for an end-to-end connection between individuals, their subjective evaluations of badges, and learning-relevant outcomes such as engagement. Qualitative studies suggest that an element of perceived relevance may be key: individuals’ assessments of whether the badges are relevant to them in terms of value, desirability, and contextual meaning could indicate their receptivity to badges’ effects. To investigate, I begin by looking for general patterns across existing programs (Chapter 2), then adopt a design-based approach in which I design a badge system that attempts to implement sound evidentiary practices (described in Chapter 3), while also collecting more detailed longitudinal data about perceived relevance of badges by individuals to investigate its potential role in badge processing (Chapter 4). In Chapter 5, I discuss the continued direction of this work, evaluate progress toward answering the overarching questions, and the implications of any findings so far.

1.1 LITERATURE REVIEW

At a mechanical level, a digital badge is remarkably simple. In response to an event and optionally some environmental variables (McDaniel & Fanfarelli, 2016; Hamari & Eranti, 2011), a computer program generates a packet of data containing an image, a few text fields, and possibly some metadata linking to other data.

Of course, to truly be considered a badge and not just badge-shaped bits, semantics must be attached to the various text fields and images. At this stage, the digital badge is still a simple object: one text field now describes an achievement that the badge is associated with, another gives it a name; the image is now understood to be a visual representation of an accomplishment; a list of criteria for earning the badge are written in, and the name of the issuer is stamped on. One of the metadata fields now points to evidence of the accomplishment having taken place. As the triggering event occurs, names are added to a list of recipients who have earned the badge – they take copies of their badges with them, and an open badge is born (Open Badges Project, 2017).

The rest, as it turns out, is on us as viewers, recipients, and designers. Symbols of accomplishment have a long and storied history in human culture, not all good (Ellis, Nunn, & Avella, 2016; Halavais, 2012; Ostashewski & Reid, 2015). In the past few years, attention has turned particularly to the use of digital badging in the world of education. Such systems draw inspiration from a number of sources – Scouting traditions like the Boy Scouts and Girl Scouts, online reputation systems like the one used to indicate high-quality posters on the community Q&A site Stack Overflow, and video game achievement systems like Xbox Achievements.

Because it is important to understand the context in which digital badges are being envisioned, I will begin this chapter by examining the literature describing the potential of Open Digital Badges, and providing a non-exhaustive summary of places in which digital badges are being used currently. I then move to the primary research topic of motivational effects of digital badging, starting with a brief overview of proposed motivational mechanisms by which badges might operate, then reviewing select empirical studies around a promising direction of modeling *motivation to earn badges*, separately from *motivation in the domain*. Finally, I present a broad research agenda including a provisional theory of action, and situate the three studies contained in Chapters 2, 3, and 4 within the course of that research.

1.1.1 The Educational Promise of Open Digital Badges

Despite their technical simplicity, digital badges may hold a great deal of potential. Devedžić & Jovanović (2015) reviewed the literature around digital badging and organized potential value propositions around four key stakeholder groups in education. In their analysis, they identified that learners might benefit from badges’ ability to support goal-setting, planning, and self-reflection; from providing and integrating feedback across environments; from recognizing skills that aren’t traditionally recognized; and by developing their sense of community membership. Teachers might expect badges to motivate and engage their students, help them scaffold learning by charting learning routes for students; and by acting as alternative assessment. K-12 schools might benefit if badges facilitated novel assessment, grading, and feedback processes within them, promote the visibility of their schools, facilitate inter-school collaboration, and jump-start instructional improvement through the process of designing them. Employers might benefit from using badges as an efficient tool to locate individuals with specific capabilities, by having badges supplement existing credentials, from badges’ potential to recognize “soft skills”, and from the visibility that might come from offering or accepting badges; however, employers are often concerned from the validity of badge credentials. And finally, other organizations in the educational ecosystem might benefit from badges recognizing extracurricular learning, promoting learning opportunities, and bolstering the reputation of municipalities that issue them.

1.1.1.1 Enriching learning ecosystems

The MacArthur Foundation’s *Connected Learning* vision posits that the norms around recognizing and developing skills and knowledge in the United States are fundamentally impoverished because they fail to recognize sources of learning other than formal schooling, even though children spend the majority of their time in places other than schools, on top of the fact that many schools are ineffective to begin with (MacArthur Foundation, n.d.).

Democratization: Badges will enrich learning ecosystems by allowing new actors to issue educational credentials. One common thread in badge literature is the idea that schools occupy something of a monopoly position in the world of educational credentialing. The underlying logic behind this claim is that there are a great many places other than school where people learn valuable skills, but that such learning is not recognized (MacArthur Foundation, n.d.). Since (open) digital badges can be issued by anyone, they hypothetically provide organizations like libraries, museums, and afterschool spaces with the means to make and substantiate claims about skills that participants have acquired outside of formal school. A side benefit of this, tied to the *micro-credentialing* claim (see Claim #2), is that small programs with tightly focused offerings can offer badges on even very specific topics. This further increases the diversity and scope of recognizable learning. To the extent that their badged credentials are judged credible, the previously unrecognized learning providers become part of the accepted educational diaspora, thus "democratizing" the process of educational credentialing. I refer to this as the *democratization* goal.

At a theoretical level, the complaint of monopoly does not seem to be about schools themselves, but rather *schooling*, a cultural pattern of expectations around what education is and ought to be. Meyer & Rowan (1977; 1978) argue that "educational bureaucracies emerge as personnel-certifying agencies in modern societies" and follow a very particular kind of structure because of expectations "institutionalized" in cultural norms. Tyack and Tobin (1994) discuss a similar phenomenon, referring to it as the "grammar of schooling", and noting that the seeming inescapability of, e.g., the Carnegie unit, has been a major impediment to school reform. The success of badges depends on whether or not they can negotiate this perilous space.

While the challenges badges will face in reaching the *democratization* goal are not exactly the same as those which have been faced in the reform of existing schools, the barriers will be similar. Badges do not need to break a school out of the mold of schooling; instead, they will need to break other organizations *into* the world of educational legitimacy. And they will need to do so without simply turning those other organizations into schools.

Playlists/Pathways: Badges will enrich learning ecosystems by exposing students to options for activities or progress. If badges succeed in realizing the democratization goal, then badges will also act as a beacon for learners, indicating where learning opportunities can be found. Where there is a badge, there is something that can be learned. Rughinis (2013) suggested that badges might be thought of as "routes through an activity system". If that system is sufficiently large, then badges or collections of badges may guide learners to learning opportunities they would never have discovered otherwise. This approach is currently being implemented by the LRNG as "Playlists" that link experiences (XPs) together to form a chain of activities and evidence that leads to a badge, which in turn is expected to unlock an opportunity for the earner (LRNG, 2017).

"Stackable" portfolios: Badges will enrich learning ecosystems by illustrating complex trajectories of learning over long periods of time. As learners accumulate badges, they will have accumulated a record of tangible achievements. Buchem (2016) points out that, like digital portfolios in general, badges can facilitate the illustration of pathways of learning over time, bridge evidence from multiple settings, facilitate a learner's agency and ownership of the learning, provide evidence, and catalyze discussion about learning. However, badges are somewhat more constrained than portfolios in that the learner has a finite set of badges to choose from, badges tend to be smaller in scope, and assessment is given a higher priority. The way in which badges and portfolios overlap also have implications for design: the badges may become part of a broader portfolio, or the badges may serve as a portfolio themselves.

1.1.1.2 Assisting with credentialing and employment

Fewer than half of US employers say that a college transcript is even “fairly helpful” in helping them to evaluate an applicant. However, 80% say that an electronic portfolio would be useful (Hart Research Associates, 2015).

Evidence: Digital Badges will assist with credentialing and employment by natively embedding evidence into the credential. Well-designed digital badges back their claims by digitally embedding evidence into themselves in the form of metadata. Casilli & Hickey (2016) argue that not only does this paradigm contrast the existing reputation-based one (e.g., the value of a degree is a function of the issuing institution’s academic reputation), but that the evidence-based paradigm could ultimately precipitate a disruptive shift in the world of educational assessment and credentialing. They do not get into the details about what kinds of evidence will be compelling, rather that the “level of evidence” (p.124) will be transformational. In order to achieve this, however, the assessment logic underlying the evidence must be “transformational”; otherwise, it could be simply “conformational” (i.e., absorbed by or reproducing the old system) or even “deformational” (harmful to learners) (Torrance, 2012).

Micro-credentialing: Digital Badges will assist with credentialing and employment by recognizing more specific and finer-grained skills than traditional credentials. Smaller-scope credentials like digital badges indicate finer-grained skills compared to traditional degrees. This, in turn, enables documentation of skills at higher frequency (Grant, 2016). In addition, badges identify candidates with skills even if they do not possess a more general credential. This is particularly relevant for adult learners who have picked up skills “on the job” (and hence without degrees) to bring them into the discussion more easily (Finkelstein, Knight, & Manning, 2013).

Digitization: Digital Badges will assist with credentialing and employment by recording learning records in a standard digital format. Because badges are digital, employers can rapidly search for applicants with a particular skill via electronic means. In more complex applications, it could allow employers to search for learners with complex combinations of qualities by interrogating large badging data sets (Pursel, Stubbs, Woong Choi, & Tietjen, 2016).

Soft Skills: Digital Badges will assist with credentialing and employment by recognizing different kinds of skills than current systems. Another type of skill that employers are interested in but cannot identify through traditional means, are the so-called “soft skills” like teamwork and problem solving (Heckman & Kautz, 2012; Whitmore & Fry, 1974) – also called 21st Century Skills and non-cognitive skills. These are among the highest-sought skills by employers (Barton, 2006). Formal schooling has had trouble recognizing and communicating these skills (Bowen & Thomas, 2014). High quality out-of-school programs, on the other hand, develop these and other positive social behaviors (Durlak & Weissberg, 2010). Digital badges provide those programs with means to communicate their value in the development of soft skills. Thus, digital badges that accurately indicate applicants who possess these skills could fill a critical gap in the credentialing world.

1.1.1.3 Improving instruction

A somewhat less commonly explored goal in digital badging work is the use of badges to directly improve instruction or learning.

Assessment: Digital Badges can play a role in improving learning through assessment. Abramovich (2016) argues that digital badges are inherently a form of assessment, as certain criteria must be met for them to be awarded. The nature of this experience could be educative (cf. Wiggins, 1998) – that is, the feedback from earning, not earning, or knowing what it takes to earn the badge could help a learner to adjust his or her behavior accordingly.

Big Data: Digital Badges can play a role in improving learning by creating large, usable data sets to inform learning analytics. Badge earning and metadata patterns on large platforms such as MOOCs can be used to discover trends, make content recommendations, and facilitate adaptive learning. It could identify trends in demand, so course designers can respond more effectively, and use data about distal outcomes to suggest possible career options to learners based on their skills and interests (Pursel, Stubbs, Woong Choi, & Tietjen, 2016).

1.1.1.4 Badges today

A report by Sheryl Grant on "Promising Practices of Open Credentials: Five Years of Progress" (2016) identifies a number of important areas where digital badging programs have taken root. There were several notable forerunners in this area – Khan Academy, which introduced badges as part of a large-scale gamification effort (see Morrison & DiSalvo, 2014); FourSquare, which used badges and gamification to motivate interaction and build a community (Gibson et al., 2013), and Stack Overflow's reputation and badging systems, which demonstrated an empirical effect on user behavior (Grant & Betts, 2013).

Today, however, digital badges have been adopted in a wide variety of contexts, and for a wide variety of purposes. Overall, early adoption has followed the broad outline of the *enriching the ecosystem* goal: recognizing and including learning from previously-unrecognized sources (*democratization*), documenting that learning (*portfolios*), and connecting across opportunities (*pathways*). The primary focal area of badges appears to be *soft skills*, though a fair number of badge systems leverage the *micro-credential* scope to recognize new skills. Many of the badge development processes have focused on ensuring that compelling *evidence* is included with badges. Some examples include:

Competency-based education (CBE). In contrast to traditional education, CBE is an approach to education which “upends the traditional paradigm of credit hours and seat-time in favor of the bundling and unbundling of skills, knowledge and abilities, wherever learning may originate” (Grant, 2016, p.8). While the notion of designing education around desired outcomes is a very old one, competency-based education in its modern form draws on Bloom’s work on assessment in the 1950s-70s, a desire to increase the relevance of skills taught in schools, and concern for serving the individualized education needs of exceptional students (Malan, 2000). The current surge in interest in CBE occurs alongside an increased need for job retraining and recertification among adults and emphasizes the recognition of learning that has occurred through work experience and from other sources (Book, 2014). Competency-based education programs are currently permitted to issue diplomas in 36 states (Grant, 2016). Competency-based credentialing via badges is also being investigated in the health care field (Beals et al., 2015).

Digital badges are a natural fit for CBE initiatives because they can communicate standardized information about a competency without getting tied up in the details of where that competency was developed (Blackburn, Porto, & Thompson, 2016). This makes since, given the shared *democratization* goals between badging and CBE. Badges are mapped onto assessments and learning content of appropriate quality both in and out of school, and they are arranged into pathways aligned with standards in academic and 21st century skills in appropriate state frameworks.

Nevertheless, one of the key challenges facing CBE badges, and indeed the field of CBE in general, is the great weight placed upon the badges for valid assessment. No longer a simple indicator, their central role in CBE propels badges into the world of high-stakes decision making. It remains to be seen what impact this will have on digital badges overall.

Out-of-School Time and Informal Education. Educational providers outside formal education offer “hands-on experiential learning… often [focused] on sophisticated 21st century skills and dispositions that are essential to social and economic mobility” (Grant, 2016). Recognizing that much of the learning students are likely to do in their lives can take place outside the formal classroom, out-of-school programs including museums, libraries, and afterschool programs have made a concerted effort to provide high-quality opportunities for learning, and many have now begun to employ digital badges as a means of recognizing that learning (Jovanovic & Devedzic, 2015).

In other cases, OST providers have partnered with formal education organizations to use badges as credentials to formalize (often for-credit) articulation agreements about learners’ achievements in college/career readiness, digital literacy, and technical skills (Davis & Singh, 2015). Finally, some programs like the Computer Science Student Network (CS2N, now the Computer Science STEM Network), which develops curriculum for mixed formal and informal K-12 use, have approached digital badges as tools toward the *motivation* goal as well (Abramovich et al., 2011; Higashi & Abramovich, 2012).

Further, in pursuit of the *pathways* and *portfolios* goals, and in line with the overall *enriching learning ecosystems* direction, badge-issuing organizations have begun banding together into badging networks. The first attempt at this occurred through the 2013 Chicago Summer of Learning effort, in which dozens or hundreds of summer out-of-school learning providers banded together and used badges to increase their collective visibility (Hurst, 2015). This practice spread to additional cities through programs like the Pittsburgh City of Learning in subsequent years (Grant, 2016). While out-of-school-time (OST) networks are not fundamentally tied to badges, digital badges were introduced as a means of collecting and displaying (and thus promoting) the learning that was taking place in the networks. This involved badged programs coming together, and new programs adopting badges in order to join in.

K-12 Schools. Schools in the formal education space have used badges to "help elevate social-emotional learning and 21st century competencies… to recognize skills, dispositions, and roles within the school environment… [and] connect to community partners like afterschool providers and employers" (Grant, 2016). Recognizing the disconnect between what is communicated through traditional formal schooling credentials and the *soft skills* demands of employers, some school districts have also begun supplementing their traditional credentials with badges for 21st century skills (Derryberry, Everhart, & Knight, 2016).

In some cases, like the Passport 2 Success program at Corona-Norco Unified School District, schools have combined digital badging with gamification elements (in this case, points tied to the earning of each badge) to offer tangible rewards in hopes of increasing student *motivation* (Moore & Edwards, 2016).

Higher Education. Perhaps in part due to controversy about the value proposition of higher education programs and credentials, universities and colleges are exploring a wide variety of approaches to digital badges (Grant, 2016).

Individual instructors have taken the opportunity to design badges into their classes to *motivate* participation by students (e.g., Chou & He, 2017). Mirroring the trend in K-12 mentioned earlier, some universities are issuing badges for *soft skills*, though also in some cases for academic skills, or to capture credit for co-curricular activities such as service learning in an online *portfolio*. Colorado State University uses badges as both organizers and indicators of completion for a number of non-credit online courses (Grant, 2016), in alignment with the *micro-credentialing* goal. Universities have also designed badges for professional development of staff, such as librarians (Bebbington, Goldfinch, & Taylor, 2016). In 2014, Concordia University Wisconsin began offering an online master’s degree that uses digital badges embedded into credit-bearing courses as the criteria for program completion (Bull, 2014).

The Education Design Lab (a nonprofit consulting company) collaborates with institutions of higher education and employers to develop frameworks and technology (including badges) that align the content and credentials of higher education with the needs of employers (Grant, 2016). A collaboration between the Foundation for California Community Colleges and New World of Work has produced a set of 10 shared badges that can be verified by either college instructors or employers (Foundation for California Community Colleges, 2017).

Workforce. Companies, professional societies, and other organizations have also adopted various versions of educational (often training) badge systems. Commercial users appear to be largely concentrated in the technology space, which is unsurprising given the digital nature of the badging innovation. Both IBM and Microsoft are using them to support their existing training networks for IT professionals using their technologies. Such badging activities are well-aligned with their parent companies’ interests of encouraging adoption and widening the trained user base for their technical products. Zappos (a subsidiary of Amazon.com) is using them for internal training and advancement (Grant, 2016). Of note, many commercial badging systems do not provide means to transport credentials outside their own platforms. Thus, while they are “digital badges”, they are not “open digital badges” in the sense that Mozilla and others use the term.

Professional organizations, which often have existing credentials and certification programs, have also begun adopting digital badges. The American Institute of CPAs offer digital certificates for pre-existing credentials (AICPA, 2017). Educause, a professional organization of higher education informational professionals, uses badges to mark participation in its own activities, and also promotes badges' use in universities (Educause, 2017). Various teacher professional development organizations such as Digital Promise, VIF International Education, and Who Built America, are using digital badges to encourage teachers to develop particular instructional skills (Grant, 2016).

Organization in the service learning sector have also adopted badges. These are programs in which participants spend a significant period, typically a year or more, in a modestly compensated position that works toward some public good, such as environmental stewardship or aiding a vulnerable population. Service learning emphasizes the learning impact this has on "service fellows" who undertake this work. The Corps Network and Service Year Alliance co-developed a framework for badges that emphasize 21st century skills and college readiness, that can be developed and assessed by site supervisors across the wide variety of contexts in which fellows serve (Grant, 2016).

1.1.2 Theoretical Impacts of Digital Badges on Motivation

We now turn to the major focus of this dissertation, which concerns the interaction of digital badges and motivation. Interestingly, one of the most pervasive claims made about digital badges is that they are motivational, but this claim is often made without a particular theoretical justification. This may be a side effect of badges' lineage through video game achievement systems, which are often assumed to be *a priori* motivating. However, scholars have unpacked this claim from several theoretical angles.

1.1.2.1 Identity

Digital Badges will impact motivation through identity-building. Badges act as social signifiers that can “assist users in building and formalizing identity in social media networks” (Gibson, Ostashewski, Flintoff, Grant, & Knight, 2015). Identity, in turn, dovetails with interest, so a learner who identifies with a domain or community is more likely to continue engaging with it (Eccles, 2009).

1.1.2.2 Goals

Digital Badges will impact motivation by making goals visible. Digital badges serve as markers for next-steps. By exposing students to options for activities or progress and helping them understand what must be done, they create a “signaled route” (Rughinis, 2013), which allows learners to use and improve their self-regulated learning skills (Pintrich & De Groot, 1990; Charleer, Klerkx, Odriozola, & Duval, 2013). If these goals are set up “just outside of comfortable reach”, they should prove additionally beneficial (Antin & Churchill, 2011).

1.1.2.3 Expectancy-Value

Digital Badges will impact motivation by increasing confidence and task value. Expectancy-Value Theory (Wigfield & Eccles, 2000) suggests that people do things that they find valuable, so long as they think they can succeed, and the cost is not too great. Badges could increase the perceived value of learning by enhancing feelings of accomplishment (attainment value), by unlocking some future reward (utility value), or if already-earned badges increase confidence in success at the next step (expectancy).

1.1.2.4 Rewards

Digital Badges will impact motivation by acting as rewards. Rewards incentivize behavior. However, the framing of digital badges as rewards has caused justified concern that they might function as extrinsic motivators that undermine intrinsic motivation in the task domain (Resnick, 2012). This appears to be a real danger, but an avoidable one (Abramovich, Schunn, & Higashi, 2013). The key, according to self-determination theory (Deci, Koestner, & Ryan, 2001; Deterding, 2011; Ryan, 1983), is to make sure that badges are not perceived as *controlling* by learners.

1.1.2.5 Gamification

Digital Badges will impact motivation because they are game-like. Gamification is “the use of video game elements … to improve user experience and user engagement in non-game services and applications” (Deterding, Sicart, Nacke, O’Hara, & Dixon, 2011). Badges are a form of gamification, since digital badges have long been used in video games. That description, however, says little or nothing about how they will serve to motivate learners. As a motivational tool, badges will more likely be expected to operate either by evoking one of the other channels above (e.g., by acting as a reward, or setting up an enticing progression of goals), although it remains possible that digital badges will evoke a sense of gamefulness (Deterding, Dixon, Khaled, & Nacke, 2011) and that the learner may shift mental frames to one of “play” – if that happens, then it might be possible to draw upon the situated motivational affordances (Deterding, 2011) of that state to leverage intrinsic motivation.

1.1.3 Empirical Research on the Motivational Effects of Digital Badges

Empirical research on the motivational effects of digital badges has been steady for several years, though at times variable in quality and consistency. Here, I narrow my discussion to qualitative and quantitative studies which attempt to draw inferences about the effects or functioning of digital badges based on collected data (i.e., are not purely descriptive, theoretical, or aspirational).

1.1.3.1 Badges as intervention

Early research on the efficacy of digital badges to produce learning or motivational outcomes often found muddy effects on both. A 2014 meta-analysis of badges in computer science education (Falkner & Falkner, 2014) concluded that – in part due to the small size and lack of strength in the research base at the time – badges might be a failure. A quasi-experimental study by Filsecker and Hickey (2014) found short-term conceptual learning gains for badged vs. non-badged groups, but multiple marginally non-significant motivational outcomes (interest and disciplinary engagement; on both proximal and distal measures), with a sample of n=106 students.

More recently, however, researchers have concluded that individual-level factors appear to be moderating the effects of the badge systems on motivational and learning outcomes. For instance, the impact of badges on engagement might vary based on students' levels of prior knowledge or motivation. Not only does this realization provide analytic leverage by accounting for the omitted variable – it also represents a shift in the core question about badges from *whether* badges work, to *when* and *where* (Hickey & Schenke, 2017), and *for whom* they work.

One of the earliest and most widely cited studies on this topic is an investigation by Abramovich, Schunn, and Higashi (2013) of middle school students using a computerized cognitive tutor program in mathematics. Their analysis focused on the relationship between achievement goals, expectancy, subjective value, prior knowledge, and badge earning. They divided the course's badges into two types: skill badges, which were only earned for demonstrating proficiency in the course content; and participation badges, which were awarded for simply using the system. Over the course of the badged activity, performance avoidance goals decreased among low-prior-knowledge students (an improvement, as performance avoidance goals are associated with counterproductive learning behaviors), but this effect was weaker among those who earned more participation badges. Meanwhile, for high-prior-knowledge students, earning more skill badges was associated with an increase in expectancy of succeeding at math tasks. The authors conclude that different types of badges are more effective with different types of learners (but that participation badges might be a bad idea in general).

This effect was replicated and extended to learning outcomes in a study by the CREATE Lab (2015) showing that badges could be deliberately designed to achieve different (mixed) impacts on learning, with the effects varying by level of mastery orientation and situational interest in the course topic (geometry). Not all outcomes were positive in this study, suggesting that for some learners, badges may have undermined existing domain motivation.

One emerging idea of interest is that *motivation to earn badges* may be a worthwhile construct unto itself. Reid, Paster, and Abramovich (2015) found that learners with high expectancy toward their course domain (English) had higher intrinsic motivation to earn the course badges, including a pronounced upward bump toward the end of the course. No such effect was observed for low-expectancy students. Fanfarelli and McDaniel quasi-experimentally

(2015) identified differential impacts of badges according to the Long-Dziuban *reactive behavior types and traits* typology (Dziuban, Moskal, & Dziuban, 2000), along with self-identified phobic, impulsive, obsessive-compulsive, and hysterical traits. The authors concluded that badge-earning may be interpreted as social approval among learners who are dependent upon such approval, and failing to earn badges when available may cause negative outcomes for them – a difference in *reason* for wanting to earn the badges that foreshadowed the increased importance of doing so for that type of learner.

This link between motivation to earn badges and course outcomes remains understudied at present. Fanfarelli and McDaniel (2017) examined a subset of their data from the 2014 study to see whether a significant correlation existed between badge-earning behavior and course grade among those who had indicated that they thought badges were important in video games, or among those who indicated that they did not think so. They found no correlation, but the null result is largely inconclusive due to the small size of the study sample (n=21 before splitting into high- and low-importance response groups). Other studies have examined only isolated portions of the question. For example, Foli, Karagory, and Kirby (2016) measure a number of Likert-scale items relevant to this concept, such as *Earning the badge made the assignment more significant to me,* but report only mean response scores and standard deviations for these items, along with suggestive comments from a follow-up interview that indicated both differences in reasons and a wide range of opinions regarding the badges' subjective impact. No quantitative links were made to measured outcomes. Garnett and Button (2018) found that learner sex and game-playing behavior predict individuals' interest in earning badges as a means of helping them prepare for class, but do not report connections between that interest and actual badge-earning or subsequent course performance.

In summary, the issue of individual differences in the effects of badges is both new and old. Individuals respond differently to badges, and these differences can “swing” or even “flip” the effects of badges between positive and negative. Badge system designs can amplify or attenuate this effect for different groups. The reasons for these differential impacts are less clear. One promising direction may be to more intentionally model the desire to earn a program’s badges. There is some evidence that this intermediary factor may prove to be an important piece in the puzzle, but it is also understudied. The specifics of the underlying mechanism are unclear at this point, but the wide variety of factors that appear to be relevant suggest that the phenomenon is both real and complex, and the consistency with which it appears in the research demands that it be accounted for in future research.

1.2 RESEARCH AGENDA

1.2.1 Badges as Process

To understand individual differentiation in badge impact, I propose that another shift in the question is necessary. By and large, the existing research treats digital badges as an intervention with an impact to be measured. More sophisticated analyses have attempted to examine when, where, and for whom badges work. However, only a small body of work attempts to address the question of *how* badges work, as a contextualized process. Specifically, there must be some form of connection between learners, badges, and outcomes such as engagement.

As obvious as this sounds, relatively little existing work truly touches on this aspect. Even a random-assignment controlled experimental study of “the effects of badges” fundamentally places the badging system inside a black box – it cannot (and often does not wish to) get at the *how* question, i.e., *how* badges effect changes in learners. The inclusion of demographic or motivational factors to determine *which* learners are most affected by badges does not change this. Without problematizing and modeling the full processes by which learners, badges, and outcomes are related, we will not understand how badges work, and we cannot design optimally effective badge systems.

1.2.1.1 Engagement with badges

The most headway on this front has been made by qualitative studies. Suhr (2014) noted some members of an online community simply dismissed its badges as undesirable because they were earned for achievements they considered trivial; Davis and Singh (2015) report on a program whose badges were seen as pointless, ignored in favor of grades, and subsequently had no effect on students' activity. Wardrip, Abramovich, Bathgate, and Kim (2016) conducted a series of qualitative interviews with sixth-grade students who engaged with an optional badging program in a school environment. They identified themes consistent with the predictions of self-determination theory and expectancy-value theory: autonomy from optional involvement, self-pacing, and self-curation of evidence; recognition of competence by making the badges non-trivial, and accepting evidence from other classes; students valued its novelty, utility, challenge, and personal connection. Unfortunately, the authors did not collect data from students who did not engage, so it is unclear which, if any, of these aspects were pivotal.

Collectively, however, a pattern emerges that suggests that an important step that is being left out of most quantitative analyses: that individuals' reception of a badge system will be dependent on their evaluation of their relationship to that system. Further, this evaluation appears central to learners' uptake of the badge system – that is, their willingness to engage with the badges. This engagement with badges could be behavioral (earning them), cognitive or metacognitive (thinking about or planning with them), or affective (wanting or being excited by them). Logically, if learners do not engage with the badges, then the badges' effects – however well or poorly designed – will be neutralized. If, as is typical, badges are designed to increase learners' engagement in course activities, then learners with more positive subjective evaluations of the badging system should engage more with the engagement-increasing badges, and thus experience increased program engagement, which in turn creates more learning and motivational effects (again, assuming a well-designed program). These same elements of domain knowledge and motivation likely contribute to learners' subsequent evaluations of the badges. Finally, this process is likely to be continuous, as learners constantly re-evaluate their relationship to the domain and its badges. This process is illustrated in Figure 1.

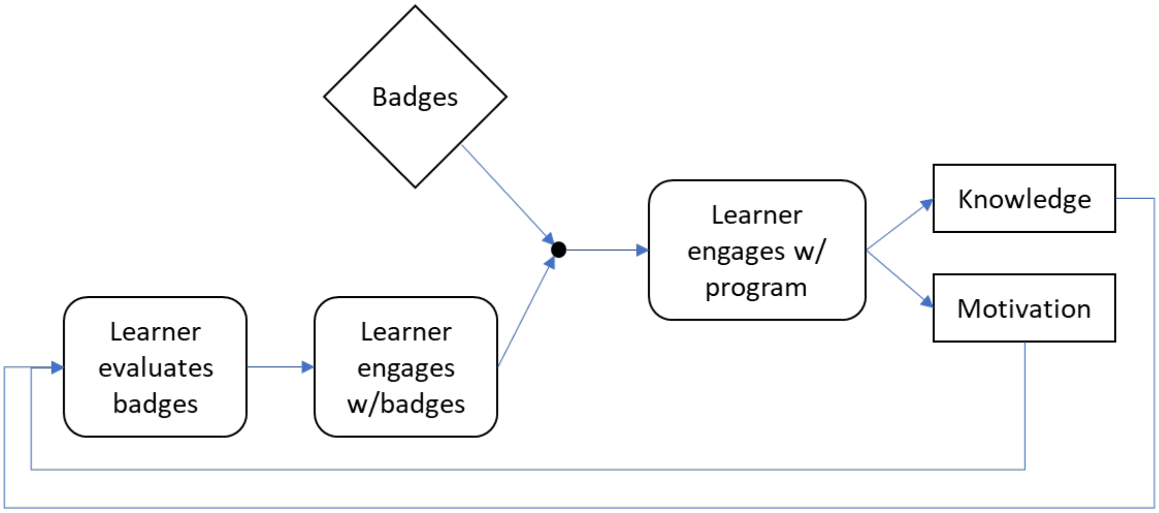


Figure 1. Theoretical model of action for learners' subjective evaluation of badges impacting program engagement and subsequent outcomes.

1.2.1.2 Perceived Badge Relevance (PBR)

This overall theoretical model suggests that there are one or more factors predicting *learners' engagement with badges* that must be modeled. As we have pointed out, recent quantitative efforts to understand individual variation in badge efficacy have begun to take up this question, but they have largely been incomplete or underpowered to interrogate this effect (e.g., Garnett & Button, 2018; Fanfarelli & McDaniel, 2017).

Again, relying on learners' reports of factors that underlie their decisions to engage or disengage from badge systems, I synthesize three initial dimensions that I believe will predict the degree to which learners engage with the badges, and therefore engage in program activities overall: believing that the badges are valuable, believing that the badge system makes sense, and wanting to earn the badges. I refer to the overall construct as "Perceived Badge Relevance" (PBR): the degree to which an individual believes a program's badges are relevant to them.

One interesting theoretical possibility that arises from this model is that there might not be any badge-specific component to the subjective evaluation at all. That is, a learner's evaluation of badges could be *entirely* contingent upon his or her beliefs about the domain and his or her individual abilities, etc. within it. This, in turn, means that the PBR hypothesis is empirically testable – if PBR explains no unique variance in program outcomes, then either the factors that learners take into account when evaluating badges are mischaracterized, or the process by which badges affect program outcomes has been mis-modeled. Both supporting and null findings will be interesting in this regard, because they will inform our understanding of how learners mentally "process" badges.

1.2.2 Overarching Research Questions

I have proposed a research direction and theoretical model which address a gap in the existing (particularly quantitative) research literature on digital badges. Through the studies presented here, I will begin to address the following research questions:

- **RQ I.** Does Perceived Badge Relevance (PBR) predict engagement in ways that support the proposed theoretical model of action? Is the predictive element unique from other motivational factors?
- **RQ II.** Do we see evidence that this process is a general mechanism, e.g., appears consistent across contexts?
- **RQ III**. Is the PBR-to-engagement process equitable, or is it biased by sex, age, or race?
- **RQ IV.** What are the design takeaways from understanding of this process?

1.2.3 Design-Based Research Methodology

Digital badges are fundamentally *designed*. Even while investigating them as a “process”, we must still look at them as a tool to produce improvements in educational outcomes and equity. Investigation into issues of design often suffers from a “minimum ontology” problem (Barab & Squire, 2004): the real-world setting of the badged program is also the smallest defensible scope at which such designs can be tested. Laboratory studies, while offering researchers great leverage to simplify and remove environmental factors, possess poor ecological validity. That is, the way learners think and act around badges in the contextual vacuum of a university laboratory may be fundamentally different from the ways they behave around real badges in real environments. Furthermore, they often lack consequential validity, as lab-grown interventions fail to take root in the real world. These criticisms are particularly pointed when the phenomenon in question is fundamentally social or contextual in nature, as they are in the case of digital badges – would Boy Scouts think their badges as valuable without seeing the buy-in of the Scouting community? Would we really be able to study badge-sharing behavior using University-owned Facebook accounts and simulated peers? Would a badge system that looks good on paper actually survive in the real world? Probably not.

This creates a difficult tradeoff for researchers. On one hand, experimental research in controllable laboratory settings are fundamentally "impoverished" with regard to context (Barab & Squire, 2004). On the other hand, real world contexts as a rule are complex, and it is very difficult to implement "clean" (or even viable) interventions there. Design experiments provide a way for researchers to conduct experiments in complex real-world settings. They involve the design and implementation of (often curricular) instruments with strong ties to the theories under study. The performance (in a broad sense, including both planned and unplanned outcomes) of the theory-aligned intervention reflects upon the validity of the theory.

As this research progresses, new data will need to be collected from ontologically valid sources. This, in turn, necessitates designing and implementing new experimental apparatus – no less than a functioning badge system with real-life users will work.

1.2.4 Overview

To test my theoretical model of badge action, I begin by looking for general patterns across existing programs in which I survey participants about PBR and engagement (Chapter 2). I then adopt a design-based research (DBR) approach in which I design a badge system that attempts to implement sound evidentiary practices (described in Chapter 3), while also collecting more detailed longitudinal data about perceived relevance of badges by individuals to investigate its potential role in the processing of badges by learners (Chapter 4). Finally, in Chapter 5, I discuss the continued direction of this work, evaluate progress toward answering the overarching questions, and summarize the implications of any findings so far.

2.0 CHAPTER 2: PERCEIVED RELEVANCE OF DIGITAL BADGES PREDICTS LEARNER ENGAGEMENT ACROSS SUMMER PROGRAMS

Digital Badges are digital artifacts that function as markers of accomplishment. Ostashewski and Reid (2015) provide a brief historical context for digital badges, noting their apparent lineage from marks of pilgrimage or political affiliation in the Middle Ages, relation to military medals, and ties to Scouting movements. Modern digital badges draw most strongly from two traditions: Scouting (as in the Boy Scouts or Girl Scouts) and video games. Typically, digital badges consist of some form of knowledge, accomplishment, or skill claim, and are backed by some form of attachment (metadata) that serves as evidence – perhaps a photo of the solar racer the student built, or a copy of the source code for the app they wrote.

Badge advocates such as the MacArthur Foundation and the Mozilla Foundation advocate that this "openness" of evidence can serve as a catalyst for a wholesale change in the world of education (MacArthur Foundation, 2011; Mozilla, 2015). Part of this change would come through disruption of traditional credentialing processes, which are typified as the granting of opaque blanket credentials (e.g., diplomas) by large, formal educational organizations (e.g., colleges). Replacing or augmenting existing processes with digital badges would *democratize* the credentialing processes by allowing a broader range of learning providers to issue recognized credentials, and also recognize more granular skills than multi-year degrees (*micro-credentialing*). Digital badges could also improve student outcomes directly by providing learners with feedback on their progress and accomplishments. Finally, badges are widely assumed to have some kind of motivational effect.

This motivational effect, however, is not yet well understood. Early studies of digital badges found mixed effects of badges on both learning and motivation (Falkner & Falkner, 2014; Filsecker & Hickey, 2014). A deeper examination of student-level characteristics suggests that a three-way interaction of learner knowledge, motivation, and badge system design may underlie the effects of digital badges – that is, certain types of badges will have certain effects on certain learners. For example, Abramovich, Schunn, and Higashi (2013) found that earning *performance* badges predicted motivational change, but only for low-performing students; *skill* badges were relevant for high performers. In short, there is evidence that digital badges can have motivational and learning effects, but these effects may not be uniform across both learners and badge system designs.

Individuals' interpretations of digital badges are likely at the heart of these differential effects. Foundational research on rewards and motivation suggest that a major distinction will involve whether learners perceive the presence of the badges as *informational* – providing them with feedback on how well they are doing – or *controlling* – used to enforce external control over the pace and nature of their work (Ryan, Mims, & Koestner, 1983; Ryan & Deci, 2000). Accordingly, Wardrip, Abramovich, Bathgate, and Kim (2016) identified support for autonomy, recognition of competence, and perceived value as key reasons given by sixth-grade students for seeking optional badges in their school. These and many other (perhaps idiosyncratic) factors may ultimately determine how students respond to badges.

2.1.1 Individuals, Badges, and Outcomes

One weakness in the current literature is that, even though it acknowledges individual differences in terms of badges' effects, it rarely connects these to individuals' beliefs about the badges themselves. That is, most studies have examined a wholesale relationship between badged conditions and mean outcomes, sometimes attending to learning general motivations or demographics. Only a small number have addressed learners' individual *beliefs about badges*, and those that do (e.g., Reid, Paster, & Abramovich, 2015) seldom examine whether "badge beliefs" connect individuals to outcomes. This is a substantial oversight, as such beliefs could play a key role in determining learners' reactions to badges.

For this study, we are interested in a single broad dimension of an individual's interpretation of a digital badge system: Perceived Badge Relevance (PBR). We hypothesize that the degree to which an individual believes a program's badges are relevant to them should predict the degree to which the badges affect them overall. This effect is analogous to the tap on a faucet: if a learner finds the badges irrelevant, the tap is shut and no "badge effect" will be present for that individual; conversely, if they believe the badges to be highly relevant, then they will be more strongly affected by them (either positively or negatively). Thus, perceived badge relevance should act as a moderator of other interpretive dimensions such as perceived controllingness. For example, a badge system seen as controlling would have a large (hypothetically negative) impact on a learner who sees those badges as personally relevant, but little impact at all on someone who sees them as controlling but personally irrelevant. Such effects may explain the varying effects of badges by motivation results in Abramovich, Schunn, and Higashi (2013) – high-prior-knowledge learners may not have regarded *participation* badges as relevant to them, while low-prior-knowledge learners may not have regarded *skill* badges as relevant to them. PBR could also apply at a more general level – individuals may have beliefs about badges overall.

A second key weakness in existing badge research is that it rarely tests effects that generalize across contexts. Most studies of badges occur within single programs, but the broader vision of digital badges involves local development of badges by a wide array of programming providers, including many informal and out-of-school time spaces. The recent meta-analysis of badges (Falkner & Falkner, 2014) only examined a coarse "main effect" of badges, and this meta-analysis was further limited to the domain of computer science education. If badges are to succeed, they will need to do so across many different programs, of many different types. Interpretation of a program's badges is highly situated within individual programs; badges that make sense and are seen as valuable within some programs would seem entirely out of place in another. Here, the construct of Perceived Badge Relevance may provide some critical analytic leverage. PBR can also be sensibly aggregated to the group level – high vs. low "average perceived relevance" of a badge system as a measure of general badge system quality and can be distinguished from individual measurements. We apply this approach to examine group-level differences in levels of perceived badge relevance, connect them to program-level factors, and even examine whether the relationships between PBR and other outcomes are similar or different between programs.

Our study thus introduces a novel construct – Perceived Badge Relevance (PBR) – and addresses two unanswered questions about digital badges:

1. Does Perceived Badge Relevance predict student engagement in program activities at an individual level, across a diverse array of learning settings?

2. Does the strength of this relationship differ substantially between programs in a way that would indicate the existence of important design or implementation differences in badge systems between those programs?

2.2 BACKGROUND

2.2.1 Engagement

Engagement, originally “school engagement” (Fredricks, 2004), captures a learner’s involvement and active participation with school. It has been extended to apply to out-of-school programs as well, as well as examined in contexts involving both types of programs (e.g., Dorph, Cannady, & Schunn, 2016; Ben-Eliyahu, Moore, Dorph, & Schunn, 2018). As a construct, engagement maintains an explicitly dual nature as a theoretically coherent whole, yet it is simultaneously composed of three qualitatively different *types* of engagement. Behavioral engagement captures learners’ choices of activities, and behavior within them. Cognitive engagement captures learners’ “mental engagement” with problems and tasks. Affective engagement encompasses feelings and attitudes that learners have toward the task. Engagement is the broader whole formed by these three parts, precipitated by the learner’s motivation. So, for instance, students who are interested in a topic (motivation) choose to participate in activities related to it (behavioral engagement), mentally engage with questions around it (cognitive engagement), and develop positive attitudes toward it (affective engagement).

For our research, we adopt a formulation of Engagement used by the Activation Lab (http://www.activationlab.org/) that focuses on engagement with a particular set of activities (rather than domain- or school-related activities in general), and ask about engagement “in the moment” (rather than as a long-term quality). This scale is appropriate for badges, which are often associated with particular activities, and questions focusing on short windows of time accommodate programs of different lengths. Past work in the Activation framework has shown a high degree of correlation between behavioral and cognitive engagement in this framing, and also a sufficient amount of overlap with affective engagement that a unidimensional “engagement” construct is psychometrically sound (Ben-Eliyahu et al., 2018).

2.2.1.1 Perceived Badge Relevance

There is already substantial evidence that badges do not impact individuals uniformly. Abramovich, Schunn, and Higashi (2013) found that high-prior-knowledge students’ expectancy of success at math tasks increased with the number of *skill* badges they earned, but low-prior-knowledge students’ did not. Low-prior-knowledge students instead experienced attenuated motivational gains (i.e., had worse motivational outcomes) with larger numbers of simple *participation* badges earned. CREATE lab researchers (2015) randomly assigned students to experience different (e.g., “mastery” vs. “performance” vs. no-badges) badge systems during a geometry learning game and found that students with different motivational characteristics made gains under different badge system designs. This type of interaction appears to extend also to how learners feel about the badges themselves – Reid, Paster, and Abramovich (2015) found that students with high expectations of succeeding in a college writing course showed a sharp increase in motivation to earn the course’s badges near the end of the semester, while students with lower expectations of success did not. Thus, even though badge systems overall may be designed to achieve certain effects, and perhaps even built in ways that would attain them, there is still a unique element at the individual level that appears to constrict or amplify these effects.

The nature of this distinction is likely complex, but theoretically centers around a three-way relationship between individuals, learning activities, and badges. In task-reward situations, Self Determination Theory researchers have found that perceptions of task-relevance and task (non-)triviality affect the ways that individuals respond to potential rewards (Ryan, Mims, & Koestner, 1983). SDT's Cognitive Evaluation sub-theory argues that the key factor underlying the observed patterns is whether individuals perceive the rewards as *informational* – providing useful feedback about their performance – or *controlling* – serving as a means to dictate their behavior. *Informational* feedback satisfies the basic psychological needs for *autonomy* and *competence* in Self-Determination Theory, while *controlling* feedback thwarts them (Ryan & Deci, 2000). In Expectancy Value Theory, the value one attributes to the attainment of something (including earning a badge) is rooted in the relevance of the achievement to the self (Wigfield & Eccles, 2000). This *attainment value*, alongside one's assessment of what the accomplishment newly enables one to do (*utility value*), intrinsic enjoyment of the task (*intrinsic value*), and counterbalanced by expected effort required (*cost*), constitute a broad category of *subjective task values* that predict individuals' actions.

We derive two insights from these positions: first, that *learner subjective perceptions of the badges* (as opposed to some objective quality of badges) will likely underlie the individual-level variation in responses; and second, that *personal relevance* will be a primary component of these perceptions. This phenomenon of personal relevance has also appeared directly in the digital badging literature, e.g., in Suhr (2014), where individuals within a community of music creators took on different roles within that community and accordingly reported that some badges were more relevant to them than others.

We propose that an individual’s level of Perceived Badge Relevance will play a role in determining the impact (or lack thereof) that a particular set of badges have on them. Further, we suspect that this relationship will function like the valve on a faucet or burner, regulating the amount of impact badges have. Insofar as badges are designed to motivate learners to engage meaningfully in a particular activity, we should expect the impact on engagement to be stronger among individuals who “buy in” to the badge system, i.e., those who perceive the badges to be more personally relevant. In a modeling sense, we can think of PBR as a moderator of any potential direct effects of the badge system on engagement. When students perceive the programs’ badges to be more relevant to them, the effects of badges in amplifying existing motivational processes will be higher; so too will any effects on engagement. When students perceive the badges to be incoherent or irrelevant to them, those badges’ effect on motivational processes will be dampened.

Finally, regarding measurement of Perceived Badge Relevance, we note several recurring themes that accompany the notion of relevance in the badging literature. Suhr’s (2014) study of badges in an online music community noted some members who simply dismissed the site’s badges as undesirable because they were earned for achievements they considered trivial (e.g., age of a user’s account). Davis and Singh (2015) interviewed students in an afterschool program and found that this program’s badges were seen as pointless, ignored in favor of grades, and subsequently had no effect on students’ activity. The findings of Wardrip, Abramovich, Bathgate, and Kim (2016) echo both the connection between non-triviality and meaning, and the importance that perceived value played in students’ appraisals of school-based badges. Synthesizing across these findings, we might expect learners’ perceptions of badges’ relevance to be indicated by the degree to which they endorse a program's badges as desirable, valuable, and contextually meaningful.

2.3 METHODS

2.3.1 Sample

Our study took place in a citywide network of summer programs in a mid-sized metropolitan area located in the eastern United States. The network was defined by a set of programs receiving grant funds from a grant making organization. The network was purposely inclusive across the city, so the set of participating programs varied in terms of program focus and structure. For instance, programs varied in length from one-hour drop-in activities to whole-summer programs; program topics were diverse, including robotics programming, nature conservation, lifeguarding, media production, and retail sales internships. The grant-making organization strongly encouraged participation in the study, so 45 different programs participated, constituting almost all programs in the network. Programs ranged in size from 2-106 learners; the median program size was 13 students.

Across these 45 programs, N=1,028 participating students contributed data for analysis: responded to brief badging and engagement-focused surveys administered at the midpoints of the summer learning programs. After removal of age outliers (most of whom appeared to be either parents or program staff that accidentally filled in the surveys), the student sample was 52% female, with ages ranging from 6–21 and a median age of 13. More than half the students came from high-poverty neighborhoods (those in which >20% of households are below the poverty line).

2.3.2 Measures

Because of the free-choice and sometimes-brief nature of informal learning programs, the surveys were designed to be low threat and very short to avoid biased and large reductions in sample size due to non-compliance of students (and providers helping to administer the surveys).

2.3.2.1 Engagement

The main dependent variable, Engagement, was measured with five items (three reverse-coded) using a 4-point Likert scale. Items were selected to span the behavioral, cognitive, and affective subdimensions of the engagement construct, and were taken from survey measures validated in science classrooms (Bathgate & Schunn, 2017). Items included, “I was bored” (r), “I felt happy”, “I felt excited”, “I was busy doing other tasks” (r), and “I talked to others about stuff not related to what we were learning” (r). Responses were presented as 1=“NO!”, 2=“no”, 3=“yes”, 4=“YES!”. Scale reliability for the 5-item engagement scale was satisfactory (Cronbach’s α = .66), especially given the multi-dimensional nature of engagement; note that psychometric analyses with larger engagement scales revealed a consistent bi-factor structure with meaningful scales at the overall engagement level and at the component affective vs. cognitive-behavioral levels (Ben-Eliyahu et al., 2018).

2.3.2.2 Perceived Badge Relevance (PBR)

The main independent variable, Perceived Badge Relevance, is a latent factor representing the degree to which learners believed the programs' badges to be relevant to them, in terms of contextual meaningfulness, value, and desirability, aspects underlying badge relevance that were highlighted in prior interview studies with badge earners (Davis & Singh, 2015; Wardrip, Abramovich, Bathgate, & Kim, 2016; Suhr, 2014). It was measured using three items on a 4-point Likert scale: "The badges in this program make sense to me", "The badges in this program are valuable", and "I wanted to earn the badges offered in this program." Response choices were 1="NO!" through 4="YES!", a Likert scale format found to have low cognitive load in younger learners, generally produce equal distance item separation in IRT analyses, and support appropriate use of means across items (Activation Lab, 2015a). The 3-item scale for Perceived Badge Relevance displayed high reliability (α = .83).

2.3.2.3 Perceived success

Learner perceptions of success are likely correlates of engagement and perceived badge relevance, and thus an important covariate. A four-item scale measuring student perceptions of success in the program was measured using four items on a 4-point Likert scale (1="NO!" to 4="YES!"). Perceived success items were, "I did a good job", "It was easy for me", "I felt successful", and "I did everything well." These items were taken from survey validated through extensive qualitative and quantitative analyses across a wide range of science classrooms (Bathgate & Schunn, 2017). Scale reliability was good (α =.74). In our analyses, we collapsed this scale to a mean of the four items in order to preserve degrees of freedom; prior Item Response Theory analyses with this instrument have shown that the gaps between Likert levels are approximately equal, and mean scores are very highly correlated with factor scores.

2.3.2.4 Age and sex

Students reported their birth date and sex on the surveys. Birth dates were used to determine respondent age as of June 1 (the start of summer).

2.3.2.5 Poverty (Socioeconomic status estimate)

Socioeconomic status (SES) indirectly captures information about family financial resources for prior informal learning experiences and social support levels that could affect motivation and engagement. Students reported the neighborhood they lived in using their common names, e.g., "Oakland"; these neighborhoods are relatively small and tend to have strongly dominant SES characteristics. These neighborhood names were matched against a table of average poverty rates per neighborhood (percent of households living below the poverty line) derived from the American Community Survey 2010 census data (University of Pittsburgh, 2012). Poverty status was then classified into one of three dummy-coded bins: low poverty (<10% of households), medium poverty (10-20%), or high poverty (>20% poverty).

2.3.2.6 Program size

Program size (total number of learners in a program, combining multiple offerings of program) was estimated through the number of survey responses received from a program. Program staff were strongly encouraged by the grantmaking organization to administer the survey to all attendees in order to allow for meaningful formative evaluation of the badge program, so response rates were likely high. This variable was heavily skewed, so a $\log_2$ transform was applied to restore normality.

2.3.3 Procedure

Surveys were administered by program staff after training by the research team, between the middle and end of each program. Single-day programs administered it at the end of their activities, week-long programs administered it on the Wednesday or Thursday of the week (to avoid graduation-day confounds), and multi-week programs administered it every other week. When more than one survey was available, we used the last-administered timepoint.

2.3.4 Analyses

We used exploratory model-building procedures to examine the relationship between PBR and learner Engagement across multiple programs. Our overall strategy was to use several analyses with different modeling assumptions to achieve convergent results regarding the strength, valence, and variability of the PBR-to-Engagement relationship across programs. All analyses were performed using the Mplus version 8 software with the multi-level add-on (Muthén & Muthén, 2017). Missing data was handled using full information maximum likelihood estimation.

We began by checking for a simple relationship of PBR with Engagement, using mean scale scores (i) in a single pool ignoring program divisions, then (ii) with variance partitioned into within-program vs. between-program components, and finally (iii) allowing for the strength of the PBR-Engagement relationship to vary between programs (i.e., a random effect). Although these analyses were conducted using maximum likelihood estimation, they are analogous to ordinary least-squares regression, fixed effect regression, and mixed effects regression with a random slope, respectively.

We continued the analysis under a structural equation modeling framework, in which PBR and Engagement were treated as latent variables measured by their indicator items, rather than collapsed to their means. This approach has the advantage of modeling both the measurement processes (of latent variables by their measured indicators) and the structural model (relationships between latent factors and covariates) simultaneously. Factors loadings and regression weights are estimated using maximum likelihood, and goodness-of-fit statistics are used to determine whether the specified model fits the data or not. In multi-level SEM models, a between-group model and a within-group model are estimated simultaneously. The between-group model estimates group-level latent intercepts of indicator variables – analogous to group means, but less biased (Asparouhov & Muthén, 2006). The within-group model captures individuals' differences from these latent intercepts. Indicators can be modeled on either level or both, as long as they have sufficient variance at that level – model estimation fails if, for instance, we attempt to model student "excitement" at the between-program level but average excitement levels are nearly the same in all programs.

We thus began the SEM analyses by examining the within-group and between-group variance of each indicator and covariate to determine which indicators could be modeled on the within- level (i.e., had sufficient variance between individuals within each group), and which could be modeled on the between- level (i.e., had sufficient variance between group averages). We followed up with a two-level confirmatory factor analysis with all usable indicators to confirm the measurement structure of our latent variables (PBR and Engagement). Indicators that had weak loadings ($\lambda<.3$) on their expected factor were removed, and the process was repeated until the model exhibited good fit. Finally, we added covariates and estimated two key models: one in which individuals' PBR predicts their Engagement equally across all programs (again analogous to a fixed-effects model); and one in which individual PBR predicts individual Engagement to a potentially different degree in each program (analogous to a random-effects model). These two models were compared to determine whether PBR's effect on Engagement was positive, negative, or neutral on average; and whether the strength of the relationship varied between programs. That is, if the random-slopes model fit the data better, the effect would be better described as varying between programs, but if the fixed-slopes model was a better (or comparable) fit, then the effect is likely similar across programs.

2.4 RESULTS

2.4.1 Sample Descriptive Statistics

Mean scores of Perceived Success and Perceived Badge Relevance were relatively high – 3.28 (SD=0.57) and 3.18 (SD= 0.71) respectively, between "yes" and "YES!" on the Likert scale. Mean engagement was slightly lower, 2.98 (SD= 0.59), just below "yes". These values suggest that overall, learners felt that badges were sensible, valuable, and desirable; that they had done well in their programs; and that they were engaged by program activities. Statistically, these values suggest that the Perceived Badge Relevance scale may be near ceiling. If so, the restricted range would mean that our results will be potentially biased toward a null finding of PBR as a predictor (due to inflated standard errors), or the analyses may underestimate the magnitude of other variables' prediction of PBR (due to artificially depressed outcome variance). In short, our results may be conservative regarding significance of PBR as a predictor or outcome.

An examination of bivariate correlations (Table 1) shows that the two primary variables of interest – Engagement and Perceived Badge Relevance – are moderately correlated and we might expect to see a significant relationship between them. Perceived Success is also correlated with both, making it an important covariate to include: does PBR appear to be related to Engagement simply by being associated with Perceived Success? Of the remaining covariates, learner age has a potentially interesting relationship with engagement – they are negatively correlated (*r* = -.25) when examined without nesting, but uncorrelated (*r* = .002) at the within-program level, suggesting that age differences *between* programs may predict engagement, while individual differences in age *within* a program may not. Program size also appears significantly correlated with Engagement; however, this correlation can be biased by the fact that larger programs, by nature, also have more data points. Therefore, while simple correlations suggest that the main effect of interest between Engagement and PBR may be present in the data, it will be important to model the data with covariates and using multi-level techniques to account for nesting.

Table 1. Pearson Correlations of Scale Scores and Covariates

	Non-nested					Within-Groups (Nested)				
	Eng	PBR	PS	Sex	Age	Eng	PBR	PS	Sex	Age
PBR	.33					.33				
PS	.38	.41				.38	.41			
Sex	.01	.02	-.04			.02	.00	-.03		
Age	-.25	-.12	-.03	-.07		.00	-.05	.03	-.06	
High SES	-.03	-.03	-.07	.05	.10	-.03	.02	-.04	-.01	.05
Low SES	-.09	.02	.08	-.10	.07	-.12	.00	.05	-.02	.08
Program Size	-.25	-.05	-.09	-.00	.02	-	-	-	-	-

Eng = Engagement, PBR = Perceived Badge Relevance, PS = Perceived Success, Sex = 1 for female. Program size is $\log_2$-transformed.

2.4.2 Regression Analyses

Our first set of analyses examined simple relationships between mean scale scores. Although these analyses involve assumptions of multivariate normality and homoscedasticity, they are useful to establish a baseline model closely equivalent to conventional ordinary least squares and mixed-effects modeling. In the simplest two-factor regression model, individual's PBR scores significantly and positively predicted their Engagement scores ($\beta = .34$, $p < .001$). This remained true, if slightly weaker, ($\beta = .14$, $p = .004$) when controlling for perceived success, age, sex, and SES through the addition of covariates into the model. R^2 = 24.9% of overall variance in Engagement was explained by this set of factors.

The next set of analyses split the variance in scale scores into within-programs variance and between-programs variance. The main advantage of this approach, which is analogous to a fixed-effect linear regression model, is that it separates program-to-program variance in mean engagement scores (e.g., some programs may simply be more exciting than others) from individual-to-individual variance within those programs. Failure to distinguish between these phenomena can result in misleading conflation of program-specific effects and general relationship trends between measured variables. The key relationship for our research questions is between individuals' perceived badge relevance scores and their engagement scores (as opposed to badge quality or program quality effects) and is thus at the within-program level. With the same set of individual covariates present in the model as before, the PBR-to-Engagement relationship among individuals within the same program was significant and positive ($\beta = .20$, $p < .001$). R^2=20.5 % of individual-level variance in Engagement was explained in this model.

Finally, we extended the model by allowing the PBR-to-Engagement relationship to vary in strength between programs. If some programs' badges had particularly strong effects on engagement, but others had no effect, or perhaps negative effects, we would observe significant variance in PBR-to-engagement slopes (positive, zero, or negative respectively). However, when we ran this model, no significant variance was observed between the PBR-to-Engagement slopes in different groups (*p* = .72). The average slope estimate was also relatively close to the previous model in which all groups shared the same PBR-to-Engagement slope (unstandardized B = .19, *p* < .001; fixed slope model unstandardized B = .16, *p* < .001).

The results from the three analyses in this section are summarized in Table 2. Together, these results suggest that, across the 45 programs in our sample, the degree to which a learner perceived the program's badges to be relevant to them was a significant positive predictor of that learner's engagement, even after controlling for age, sex, socioeconomic status, feelings of success at the program's activities, and general badge quality effects on Engagement between programs.

Table 2. Summary of Simple and Nested Regression Model Results Predicting Engagement Scores

Predictors	Simple Model (n=688)	Simple + Covariates (n=988)	Two-Level Model (n=988)	Random Slopes Model (n=988)
Perceived Badge Relevance score	.215***	.184***	.199***	.190***
Variance in PBR score slope	–	–	–	.004
Perceived Success score	–	.307***	.303***	.278***
Age	–	-.223***	.019	.008
Sex	–	-.009	.024	.029
Low SES (vs. Mid)	–	-.108*	-.182***	-.193***
High SES (vs. Mid)	–	-.016	-.119*	-.166*
Engagement variance explained (R^2)	11.5%	24.9%	20.5% of within-program	N/A

Table 2 continued

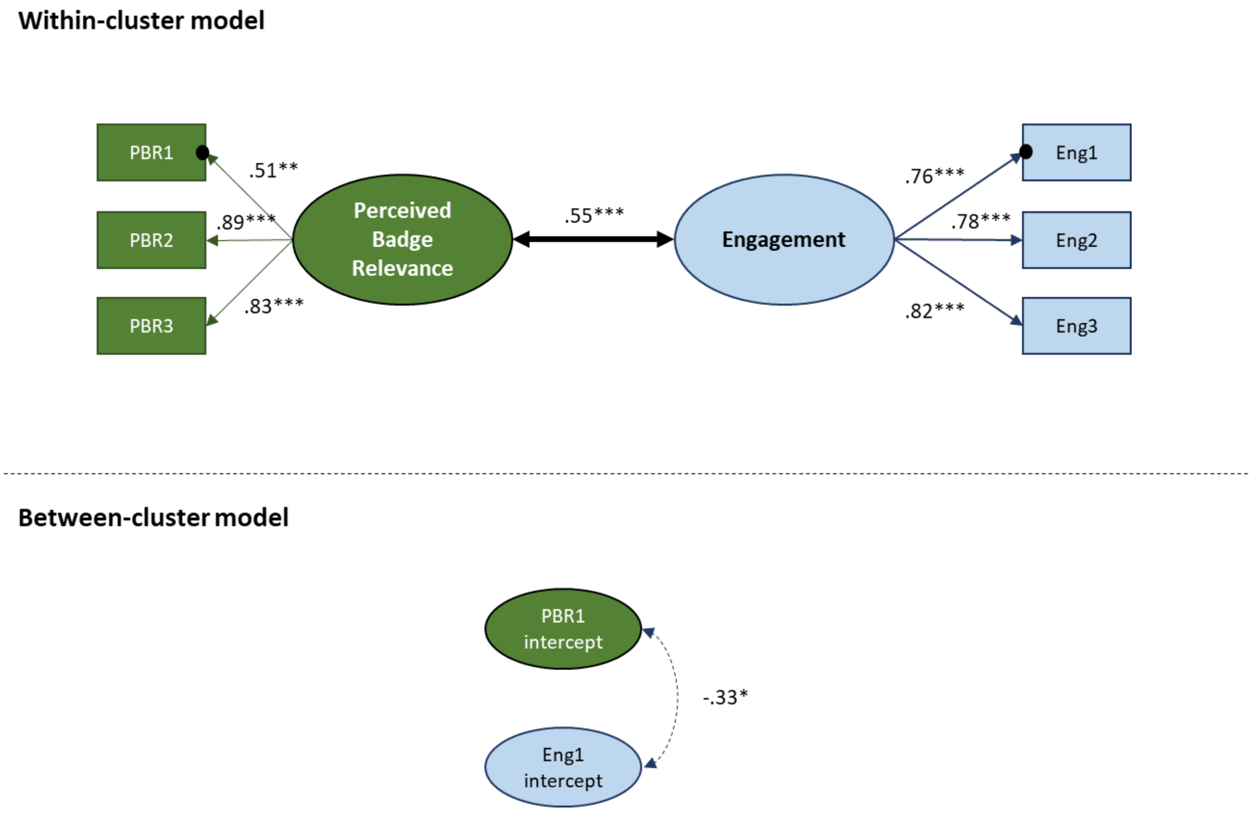
Standardized regression coefficients are shown, except for the random-slopes model. *** $p < .001$, ** $p < .01$, * $p < .05$, ~ $p < .10$. Sample sizes varies between analyses due to missingness among different included factors.

2.4.3 Structural Equation Models

Our initial set of analyses relied on a simplifying assumption that the arithmetic mean of a scale's items adequately capture variance in the construct itself. In this phase of the analysis, we use multi-level structural equation modeling techniques to both test these measurement assumptions, and to enable construction of finer-grained models with more complex relationships between factors.

This process begins by examining which measured items have sufficient variance to be modeled at the within-program (i.e., between-individuals) and between-programs levels. In particular, we examined the intraclass correlation (ICC) and variance of each item, and dropped items from the within- or between-programs level if that item fell below a certain threshold on either measure. Many of the survey items had very little variance (var < 0.1 on a 4-point Likert scale) and very low intraclass correlations (ICC < 0.1) at the between-programs level, and were thus modeled on the within-program level only. These were the Engagement items *"I felt happy"*, *"I felt excited"*, *"I was busy doing other tasks"*, and *"I talked to others about stuff not related to what we were learning"*; the PBR items *"I think the badges in this program are valuable"* and *"I wanted to earn the badges offered in this program"*; and the perceived success scale. Note that this does not mean that these items were not important; it simply means that there were not substantial differences in group means for those items. All survey items were retained at the within-groups (between-individuals) level. Thus, of the key measurement items, the Engagement item *"I felt bored"* and the PBR item *"The badges in this program made sense* *to me"* were modeled as having both between-groups differences in their latent group intercepts (roughly, their group-means), and within-groups deviations from those intercepts per individual. Among covariates, age was retained at both levels, perceived success and sex were retained only on the within-program level, and program size (a program-level covariate) was retained at the between-programs level.

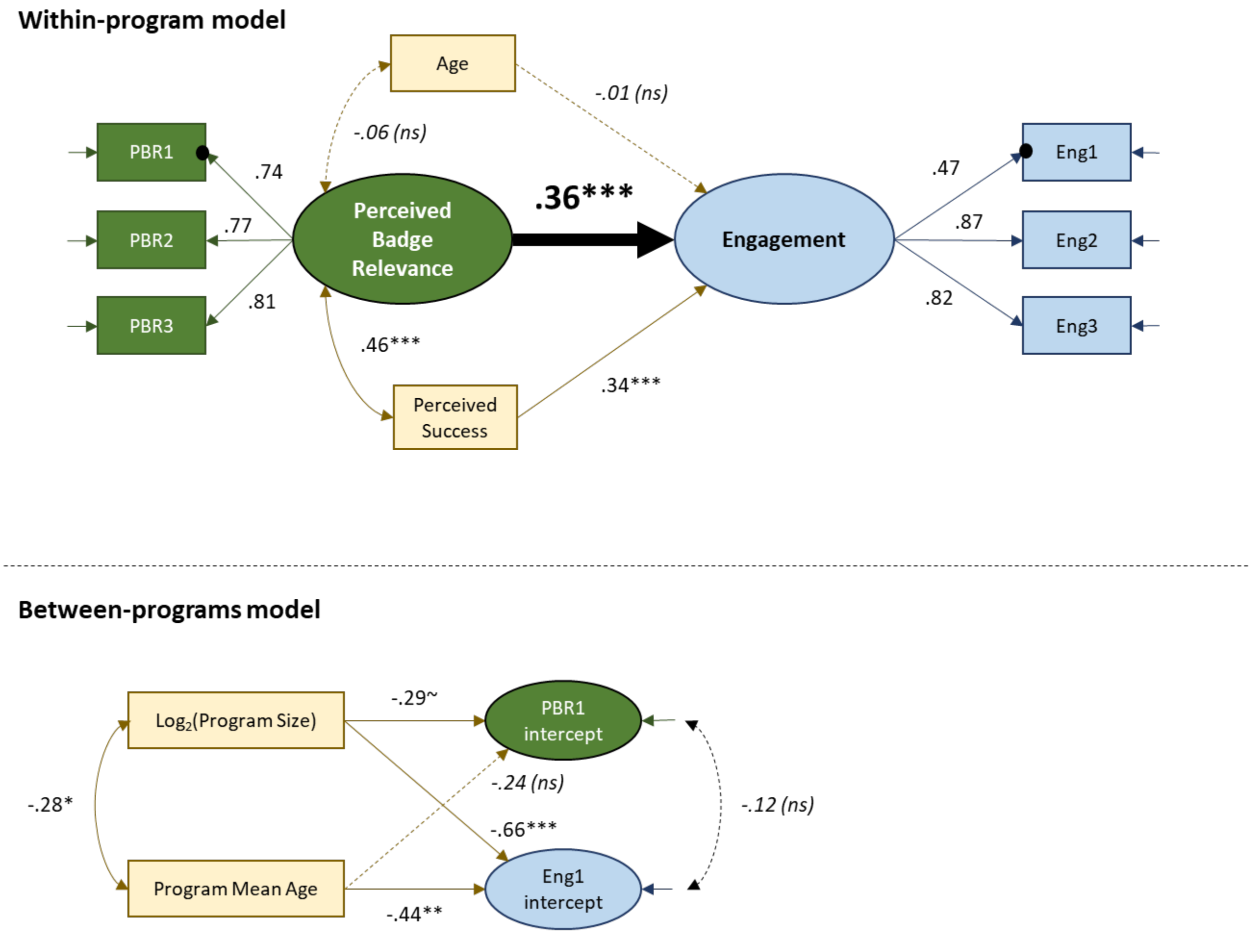
With the retained survey items, we then conducted a two-level confirmatory factor analysis. Two items in the Engagement factor had low factor loadings and were removed from the model: "I was busy doing other tasks" (λ=.14), and "I talked to others about stuff not related to what we were learning" (λ=.15). The resulting measurement model had good fit according to conventional thresholds (Hu & Bentler, 1999): RMSEA = .051 ("good fit" < .06), CFI = .975 (> .95), SRMR within = .017 (< .06), SRMR between = .012 (< .06). We used the three-item Engagement scale for all remaining analyses, noting that the retained items ("I felt bored", "I felt happy", and "I felt excited") were from the affective portion of the engagement construct, and so the latent factor might now be better interpreted as "affective engagement". No covariates were included in this step. The item-to-scale factor loadings and scale correlations are shown in Figure 1. Interestingly, PBR and Engagement were positively related within-program but the significantly varying aspects of PBR and Engagement were negatively related between programs, highlighting the importance of the multi-level analysis approach.


Within-cluster model

Between-cluster model

Figure 2. Two-level confirmatory factor analysis model after constraining indicators to levels on which they had substantial variance and removing two Engagement indicators with factor loadings < 0.3. All remaining loadings are significant at the *p* < .01 level.

Having confirmed that our measures of the latent Engagement and PBR constructs were sound, we added covariates and tested the "fixed slope" version of the predictive model, shown in Figure 2, in which PBR predicts Engagement at a fixed rate across all programs. Covariates that were non-significant on both levels were removed in order to preserve degrees of freedom and aid in model identification; the model was re-run and checked after each removal. Sex and SES were dropped at this step. The final model retained only age (both levels), perceived success (within-programs level), and program size (between-programs level) as covariates. Model fit was good: RMSEA = .024, CFI = .984, SRMR within = .020, SRMR between = .003.



Within-program model

Between-programs model

Figure 3. Fixed-slope two-level structural equation model showing standardized factor loadings and regression coefficients.

Dotted lines indicate non-significant links. The relationship of interest is bolded. *** p < .001, ** p < .01, * p < .05, ~ p < .10, *(ns)* = not significant.

In this model, individual PBR was a significant positive predictor of individual Engagement (β = .36, p < .001), controlling for individual perceived success and age, as well as group-level differences in program size and mean age. The model explains (pseudo-)R^2=37.0% of variance in the latent Engagement factor. Perceived success was also a significant predictor of Engagement (β = .34, p < .001), but age deviation from the program mean was not (p = .68).

Perceived success also correlated with PBR (*r* = .46, *p* < .001), while age again did not (*p* = .20). There were also significant relationships between program size, age, and the significantly varying aspects of PBR and engagement at the program level, again highlighting the importance of the multi-level analysis approach.

Finally, we tested the same model, allowing the slope of the PBR-to-Engagement relationship to vary between programs. Since goodness-of-fit statistics for random-slopes models are not yet well agreed upon by researchers, we used the information criteria for the fixed vs. random slopes models to determine which model was a closer and more parsimonious fit to the observed data. The Akaike Information Criterion (AIC) and Sample-Size Adjusted Bayesian Information Criterion (SSA-BIC) are diagnostic representations of the amount of variance explained by the model versus the complexity of the model (e.g., number of parameters being estimated). Thus, models that fit the data poorly or are not parsimonious are judged as “worse” according to these criteria. There are no absolute thresholds for goodness-of-fit with information criteria; however, comparisons can be made between models, with smaller numbers being preferred. For the random-slopes model, AIC =14738.435 and SSA-BIC = 14794.731. For the fixed-slope model, AIC = 12430.214 and SSA-BIC = 12488.270. Thus, the fixed-slope model is a more parsimonious fit to the data than the random-slope model. This is further corroborated by the non-significance of the PBR-to-Engagement slope variance even within the random-slopes model (*p* = .88). Therefore, we retained the fixed-slope model as our final model.

2.5 DISCUSSION

Our first research question was whether Perceived Badge Relevance predicted an individual's engagement in a badged summer program. In all analyses, Perceived Badge Relevance was a significant predictor of Engagement in the program's activities. This effect remained robust across when modeling Engagement and PBR as mean scores, or when modeling them as latent variables from the shared variance of their indicators. The inclusion of individual-level covariates, particularly Perceived Success, explained additional variance in student Engagement, but did not "explain away" the relationship between PBR and Engagement. This means that even after accounting for learner demographics, mean differences between programs, and individuals' perceptions of success at program activities, learners who perceived their program's badges to be more sensible, valuable, and desirable were more engaged in that program's activities. This also establishes an empirical baseline for digital badging effects, that they are overall positively associated with engagement (the PBR-Engagement slope would have been negative, had PBR been moderating a negative badging effect).

Our data set allowed us to use two-level models to separate *differences between programs* from *differences between individuals* within those programs. The relationship between participant age and engagement is primarily a between-programs effect: programs that serve older learners report lower engagement overall, but older and younger learners in the same programs are equally engaged. Item-level analyses reveal that more specifically, participants in programs serving older learners reported being more bored. Similarly, learners in larger programs also reported being more bored. Program size and program mean age were negatively correlated with each other in our sample (programs for younger students also tended to be larger), so in analyses that did not explicitly model them simultaneously, these two effects would have masked each other.

Multi-level modeling also allowed us to test our second hypothesis, that the relationship between PBR and Engagement would vary significantly between programs. In fact, both the scale-score and the latent-factor (SEM) analyses suggested that there was no significant variance in slopes between programs. That is, regardless of the program content, badge system, or implementation, individuals' perceptions of a program's badges as relevant to them had an overall stable, positive relationship to their engagement in the program. Thus, the observed relationship between PBR and engagement is consistent with one in which positive perceptions of a program's badges generally drive (or accompany) ongoing engagement. We found no evidence that differences in badge implementation across programs resulted in stronger/weaker (or negative) PBR-to-Engagement relationships.

Note that this does not mean that all badge systems were equally effective – there was still variance in mean PBR levels by program, which means that some programs' badges lent themselves to higher PBR ratings. This program mean variance occurred almost exclusively in one survey item, rather than across all three items of the PBR construct. Among badge-related items, the item *"The badges in this program make sense to me"* varied significantly among programs, while *"The badges in this program are valuable"* and *"I wanted to earn the badges offered in this program"* items did not. This suggests that while students in different programs felt similarly about the value and desirability of their programs' badges, they did not find all programs' badges equally sensible. This is not because younger students were unable to grasp the concept of badges – the relationship of program mean age to badge sensibility was non-significant. Instead, it may point to an endogenous issue of badge system design and implementation – perhaps the badges in some programs were confusing, or perhaps they seemed out of place in certain programs. All three items nonetheless had variance at the individual level, and only moderate covariance. This suggests that individual students varied in their appraisal of badges' overall relevance to them – and this overall level of personal buy-in was predictive of their engagement in the program – but that the three items captured meaningfully different ways in which individuals found the badges relevant.

2.5.1 Theoretical Implications

Our study makes three key contributions to the research literature on digital badging. First, our study models the relationships of students' attitudes *toward the digital badges themselves*. Most of the research around badges has not directly measured student attitudes toward badges, instead modeling badges as a black-box intervention that affects academic outcomes or domain motivation. In the course of this study, we proposed and gathered initial evidence in support of an operational construct we termed Perceived Badge Relevance, which we theorized as a moderator of the effects of badge systems on individuals. Like previous researchers who looked at attitudes toward badges (e.g., Reid, Paster, & Abramovich, 2015), we found evidence of a relationship between learners' beliefs about their success in a course, and their attitudes toward its badges. In this case, they jointly predict individual engagement, across multiple programs.

The second theoretically interesting finding is that this predictive relationship of perceived badge relevance to engagement appears to be stable across programs – across the 45 programs in our sample, no significant variance was detected. In such a diverse badging ecosystem, one would expect badge systems of different designs to have varying levels of efficacy, which would then be gated by individual learners' "buy-in" to those badge systems

(PBR). Such a hypothesis predicts that we should see PBR's relationship with engagement vary in response to how "strong" each program's badge system is. Yet no such variance was found. One interpretation is that all the badge systems in our sample were of roughly the same quality, and thus individual "buy-in" (PBR) titrated this effect at a constant rate. A second potential interpretation is that, rather than acting as a *moderator* for varying effects of different badge system designs, perceived badge relevance may in fact be a *mediator*. That is, badge systems may impact individuals' engagement *by means of* their beliefs about the badges, with "better" or "worse" badge systems simply producing more or less of these beliefs, accordingly.

Our final group of theoretical contributions stems from the fact that our data set allowed us to separate individual-level variance from group-level variance. Perceived badge relevance strongly predicts individual engagement, and individuals vary on their responses to all three measurement items; however, only "badges made sense" ratings varied between programs – learners' average appraisals of the badges' value and desirability did not. Similarly, student age within programs did not predict either PBR or engagement; however, average program age had a marginal negative predictive effect on average ratings that the badges "made sense" – this suggests that learners across the observed age range (middle and high school) are not reacting differently to badges because of their age, but rather, badges in programs serving older learners (which may have other differences in demographics) are being systematically designed in ways that are less easy to understand, or do not fit the programs well.

2.5.2 Implications for Practice

Instructors in badged programs may not have control over the design of a badge system. However, our analysis shows that, even within the same badge system, individuals who think a program’s badges are sensible, valuable, and desired – that is, whether the badges are relevant to them – are more engaged in program activities. Thus, while our analysis cannot say that badges are *causing* higher levels of engagement, it does suggest that on-the-ground efforts to ensure students feel the badges are sensible, valuable, and desirable could be positive for the overall program experience. Of note, students who feel positively about the badges also tend to be the learners who see themselves as succeeding at program activities. These two factors are both significant simultaneously – that is, the benefit of seeing the badges as relevant, and of seeing oneself as successful “stack”; each appears to be good, but both appears to be better. Program staff should continue to support both badges and other aspects of student motivation at the same time.

In terms of badge development, it behooves designers to pay attention to which aspects of perceived badge relevance show evidence of being malleable between programs. Learners in all programs felt roughly the same on whether their programs’ badges were valuable or desirable, which may suggest that badge system design does not influence those factors much. However, different programs received higher or lower average ratings on whether their badges made sense. Therefore, badge system designers may want to focus on making sure their badge systems are comprehensible, coherent, and appropriate for their environments, rather than emphasizing their value or importance to students.

2.5.3 Limitations and Future Directions

This study represents an initial exploratory step toward the unpacking of badges’ motivational effects in out-of-school programs. The single largest limitation of the study is that it is cross-sectional in nature. This prevents us, for instance, from fully addressing the question of whether or not some badge designs may cause “undermining” of learner motivation by replacing intrinsic motivation with extrinsic motivation (CREATE, 2015; Resnick, 2012) or by causing the badges to be seen as controlling rather than informational (Deci & Ryan, 2000). A second limitation concerns the lack of data for individual-level motivational factors which could function as a common cause of both perceived badge relevance and program engagement. Finally, because all the programs in our cohort were part of a single network, and received the same training, our findings may not generalize to all types of badge systems.

3.0 CHAPTER 3: COORDINATING EVIDENCE ACROSS LEARNING MODULES USING DIGITAL BADGES[1]

3.1 INTRODUCTION

No matter how successful a learning module or intervention such as an intelligent tutoring system is at producing learning, the fruits of those efforts cannot be employed efficiently without a suitable means for representing and conveying which learners possess which skills. Who will know to hire or promote this more knowledgeable individual, if there is no clear sign that he or she is more accomplished? Digital Badges are digital artifacts that function as markers of achievement. Often described as building on the combined traditions found within Scouting (e.g., Boy Scouts or Girl Scouts) and online gaming (e.g., Xbox Live Achievements, Playstation Network Trophies), badges are issued to an individual when the individual meets specific criteria embedded in program-relevant activities (Ostashewski & Reid, 2015).

Functionally, badges are commonly framed as open digital microcredentials (e.g., Ifenthaler, Bellin-Mularski, & Mah, 2016; University of Minnesota, 2017). Openness means that any party should be capable of issuing badges. Digital means that the badges themselves exist in an online environment and are thus amenable to digital transmission, e.g., over the internet. And

[1] Ross Higashi and Christian Schunn, University of Pittsburgh; Vu Nguyen, Carnegie Mellon University, & Scott J. Ososky, U.S. Army Research Laboratory.

finally, the “microcredential” nomenclature emphasizes badges’ common purpose with traditional credentials such as diplomas and trade certifications, but with a finer grain size.

A 2014 survey of an early-adopter cohort of badge developers identified three common design goals for badges: “Recognizing Learning”, “Assessing Learning”, and “Motivating Learning” (O’Byrne, Schenke, Willis, & Hickey, 2015). A growing number of studies (e.g., Abramovich, Schunn, & Higashi, 2013; Reid, Paster, & Abramovich, 2015; Suhr, 2014) have investigated effectiveness in individual areas, but the juxtaposition of the three is also informative. This is because, as with existing credentials, the assessment, recognition, and motivational components of badges are intertwined: the *recognition* afforded by a credential depends upon the fair *assessment* of the skill during the awarding process. Earners may be *motivated* to gain the credential because it is instrumental for scholastic or career advancement (utility), because it displays their prowess to others (achievement), or perhaps because they see owning the badge as consonant with their personal or professional identities. In all cases, the link between possessing the skillset and acquiring the badge depends, fundamentally, upon the validity and credibility of the assessment process. An attentive evaluator would reject (with prejudice) a badge that claims one thing but measures something else, and no learner would be excited to earn a token thus discredited.

Thus, digital badges could substantially improve the efficiency of skill-based personnel or resource assignment by effectively surfacing learners’ skills as they are developed, at a higher frequency and with greater granularity that traditional credentialing processes, although still at grain sizes large enough to be meaningful to outsiders. This may bring with it advantages for learner motivation, and by extension, improve learning outcomes as well. Yet, these things are only possible if the badges contain a valid and credible assessment of the indicated skills. In essence, the entire badging enterprise – and indeed, that of micro-credentialing in general – hinges upon the question of why a viewer should believe the badge's claim.

In this chapter, we present a conceptual model for a badge system, illustrated within a computer programming learning environment. The model is built upon theoretical foundations and practical use cases, which are leveraged in order to derive specific design considerations. The chapter concludes with the potential expansion of the badge system, and opportunities for future research.

3.2 RELATED RESEARCH: DESIGNING BADGES FOR ASSESSMENT

3.2.1 Theoretical Framework

In order to productively connect assessment and evidence, we turn to the Evidence Centered Assessment Design (ECD) framework laid out by Mislevy, Steinberg, and Almond (2003; Mislevy, 2006). Under this framework, an assessment is fundamentally understood to be an *argument from evidence*, designed for a purpose. An assessment is valid if (and only if) its argument is sound, using observed data – things the student has done – to warrant claims that the student knows certain things. Furthermore, the knowledge claim must be useful toward a real *purpose*, i.e., relevant to a real decision about a student possessing a skill. In short, a valid assessment makes a claim that an individual knows something, backs that claim with evidence, and leads to a conclusion that is usable for a decision.

How, then, might we design digital badges to embody a valid assessment claim? In the process of design, it is typical to connect these elements in reverse, reasoning from ends to means in a “backward design” process (Wiggins & McTighe, 2011). Intuitively, *purpose* will determine “which features and expectations are central, and which are irrelevant” (Messick, 1994; in Mislevy, Steinberg, & Almond, 2003). This means that we must start with the *purpose* of the badge, i.e., why anyone cares whether a student possesses a certain skill in the first place. From there, we can ensure that the badge makes an appropriate claim to fulfill that purpose, and supplies evidence that compellingly backs its claim.

3.2.2 Use Cases

In order to gain better traction on this issue, we will focus on three specific use cases. “Use cases” identify specific, representative scenarios in which the product-under-design must fulfill a certain need, in a certain context. These concrete scenarios allow designers to understand requirements and reference the scenarios as litmus tests for the sufficiency of proposed designs.

Much of the policy interest in badges frames them as credible indicators of knowledge or skills, usable for making decisions about admission to program of study or employment, or for guiding one’s own learning (Duncan, 2011; LRNG; MacArthur, 2011). Therefore, we will begin by proposing the following three use cases:

Use case 1: Digital badges should help a **college admissions officer** decide whether an applicant possesses *sufficient academic preparation to begin learning college-level content*.

Use case 2: Digital badges should help a **learner, mentor, or intelligent tutoring system** *choose an appropriate next task* or topic for learning.

Use case 3: Digital badges should help an **employer** decide whether an applicant will *be able to perform certain tasks well on the job*.

3.2.2.1 Aligning purposes, claims, and arguments

Across the three use cases, two distinct kinds of “purposes” have appeared.

Claims about Readiness to Learn. The college admissions officer and student/mentor (use cases 1 and 2) are both interested in the learner’s readiness to learn certain new content. The logic implicit in this framing is well-established within the learning sciences: learners can only learn certain novel concepts after certain prior learning has put them within reach of it. This type of claim is typically most relevant in cases of formative assessment – that which is intended to inform mid-course adjustments in learning trajectories.

Research related to the Zone of Proximal Development (Chaiklin, 2003; Vygotsky 1933), Conceptual Change (DiSessa & Sherin, 1998), Learning Hierarchies (Duncan & Hmelo, 2009), and effective Tutoring (Koedinger, Corbett, & Perfetti, 2012; Wood, Bruner, & Ross, 1976) has unpacked theoretical and practical concerns around this phenomenon. For our purposes, this distinction is particularly important because the readiness-to-learn *purpose* informs the *kind of assessment argument* that is needed to support it. Specifically, the argument must allow us to conclude that the student possesses knowledge *X* in such a way that it has prepared them to learn *X’*.

Claims about Proficient Reapplication. The importance of alignment is also evident when we consider the second major *purpose* contained in our use cases: that of the hiring manager (use case 3), who may be primarily interested in whether a job applicant will be able to perform certain skilled job functions reliably once hired. That is, the hiring manager wants to know that the applicant will be able to apply the skill proficiently and appropriately under working conditions. This type of claim is commonly associated with summative assessment – that which is intended to summarize qualifications.

The logic underlying this framing is of a wholly different nature: it concerns the *transfer* of learning from the context in which it was learned (e.g., through a lecture in the classroom or in a training environment) to new contexts in which it should be applied (e.g., to a problem-solving task in the workplace). Learning scientists have given extensive attention to the fact that two students who appear to understand something well, may still differ in their ability to "transfer" that knowledge to a new situation. This is often a concern for intelligent tutors or simulation environments in which the learning environment is systematically simplified in order to make it digital. Many models have been proposed and tested of both the underlying causes and ways in which transfer performance might be improved (e.g., Hammer, Elby, Scherr, & Redish 2005; Kirschner, Sweller, & Clark, 2006). An assessment argument for this purpose will likewise need to be of an entirely different character from a readiness-to-learn argument. Rather than focusing on a student's ability to comprehend future material, this purpose demands that the assessment argument be made about the student's ability to transfer knowledge to the work context.

3.2.3 Design Takeaways

In practical terms, our digital badge designs must make the assessment claims that speak to both the "readiness to learn" and "reapplication" *purposes*, and back those claims with evidence. Stated in "forward" order, digital badges make assessment claims of the form: "Using past performances as evidence, we assert that the earner of this badge possesses the indicated skill, and will be able to apply it appropriately and build upon it in the future".

With these general design objectives in place around the alignment of *assessment claims* to *assessment purposes*, we now turn to the specifics of the "evidence subsystem" that facilitates the formation and delivery of an appropriately "backed" assessment argument.

3.3 DESIGN PRINCIPLES FOR A BADGE-BASED EVIDENCE SUBSYSTEM

In this section, we propose detailed principles for the design of *evidence subsystems* within broader badging arrangements, designed to support valid arguments-from-evidence about the knowledge and skills of badge earners. Our objectives are twofold: (1) to more specifically address concerns about *establishing evidentiary warrant*; and (2) to provide one concrete, practical solution.

3.3.1 Design Principles

As a design progresses from “clarified problem statement” to “specific solution”, it sometimes acquires features which might be considered idiosyncratic and do not further any design goals. To make a small but powerful set of implementation decisions, while avoiding superfluous prescriptions altogether, we will rely on a small set of guiding principles to help us maintain design discipline around the complex notion of “evidentiary warrant with fidelity to purpose”. These principles reflect, in a sense, the two complementary facets of a parsimonious design: 1) the provided evidence must be *sufficient* to establish (or “warrant”) the claim, yet 2) are no more complex than *necessary* due to practical and logistic concerns. In both cases, we will draw again on our use cases, as both necessity and sufficiency are defined relative to *purpose*.

3.3.1.1 Principle 1: Evidentiary strength

Strength of evidentiary warrant increases with both quantity and diversity of evidence. To address the issue of sufficiency of evidence, we will draw upon the intuition of the replication study, or triangulation. In the sciences, we recognize that no single experiment, study, or perhaps even theory provides a complete picture. The single datum is thought of as fundamentally impoverished, lacking a robustness of perspective that can only be established through consideration from multiple angles. A conclusion supported by only a single data collection and analysis is at best promising, but provisional. Analogously, an assessment statement is only weakly warranted by a single piece of evidence.

Furthermore, while additional evidence of the same type would increase our certainty in the conclusion somewhat, it does not achieve the same effect as a concurrent result from a fundamentally different analysis since the inherent weakness of any single source is maintained in an exact replication. Real replication is not duplication of an analysis, but the reproduction of an equivalent result in a different context.

We extrapolate three important design features from this principle. First, we recognize that evidentiary warrant is not dichotomous, but dimensional — a small amount of evidence would provide weak support for the assessment claim, while more or better evidence would provide stronger support. It is not all-or-none. Second, we recognize the existence of multiple, qualitatively different types of evidence that might speak to the validity of the skill or knowledge claim. This is concurrent with a central theme of modern learning science research: in addition to the classical cognitive theories of conceptual understanding (typically assessed by, e.g., standardized multiple-choice exam), several strands of research focus on environmental and social factors involved in promoting the development of transferrable skills (e.g., Gresalfi, 2009; Lave & Wenger, 1991; Lee, 2008). And finally, as a combination of the first two, we recognize that the best conceptualization of the evidence space is in fact multidimensional: different amounts of different kinds of evidence.

We conclude with two design decisions based on the design features above.

Design Decision 1: The framework must support the inclusion of multiple kinds of evidence concurrently. The badge system must, at a technical level, support the inclusion of evidence types beyond traditional exam scores. For instance, a portfolio of works completed by the student may be considered a valid form of triangulating evidence. In the Evidence-Centered Assessment Design framework, each piece of evidence is free to rely on its own theoretical sense of internal validity, as long as it is valid according to that standard.

Design Decision 2: The design must recognize and represent multiple "levels" of evidentiary warrant. These should be tied to the quantity and diversity of evidence provided. Strength of warrant must be kept meaningfully separate from the level of mastery being claimed.

3.3.1.2 Principle 2: Evidentiary necessity

The strength of evidence needed to support a claim is based on the weight of the decision that will be made. Stronger warrant is needed for higher-stakes decisions. Having established that the strength of evidentiary warrant can vary with the amount and diversity of the evidence, the second issue we must address is, "How much evidence is enough?" Since there is no definitive scientific answer to this question (Popper, 2005), we borrow an intuition from legal theory, where the use of imperfect evidence to justify conclusions is common practice. This intuition is that of the sliding "standard of proof", in which the required strength of evidence is higher when the potential consequences are greater (e.g., in criminal versus civil court).

While the circumstances of a badge evidence evaluation are certainly not the same as being in court, decisions made using badges do vary in terms of potential impact. A student's decision to move on to the next chapter (based on a badge saying they understood the previous concept) is fairly low-stakes, and can be made based on even relatively thin evidence because even in the worst case, a learner need only backtrack to review. On the other hand, a college admissions officer is making a substantially higher-stakes decision when he or she uses badges to determine an applicant's readiness to learn in college, and should be bound to accept only evidence that is more firmly established. Thus, we extrapolate one important design feature from our consideration of evidentiary necessity: our design must indicate in some meaningful way the strength of evidentiary warrant it provides, so that a viewer can have some sense of how firmly its claim should be considered established.

Addendum to Design Decision 2 ("2b"): The design must state or strongly suggest a clear relationship between stakes-appropriateness and strength of evidentiary warrant. This is functionally an addition to Design Decision 2 from the previous section.

3.3.1.3 Principle 3: Efficiency

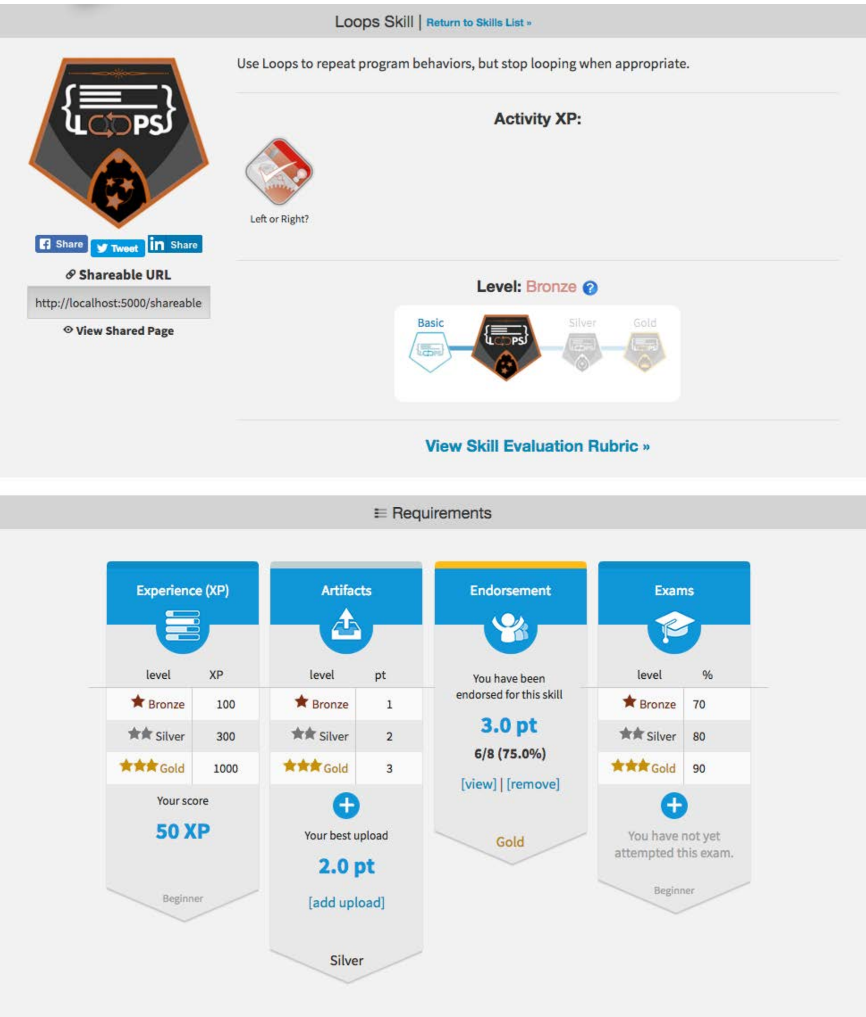
There is a practical upper bound on how much evidence an evaluator is able and willing to examine. The more efficient the presentation, the more evidence can be used. Finally, we look at a practical factor which has implications for how much evidence we can effectively (rather than theoretically) bring to bear in backing our assessment argument. Simply put, no evaluator has time to interrogate all the evidence in detail, for every badge that is presented. Looking back to our use cases once more, it is the highest-stakes decision makers — the college admissions officer and the potential employer — who are also tasked with evaluating the largest quantity of badged claims. Yet, per our second principle, these are the decision-makers who must consider the evidentiary claims most carefully. Thankfully, good communication design techniques can mitigate this information bottleneck by distilling complex and ungainly information into easier-to-understand visual summaries that can be easily reviewed without loss of the "big picture" regarding evidentiary warrant.

Design Decision 3: Each type of evidence in the badge must be summarizable for quick viewing. The composite strength of the badged evidence claim should also be easily summarized.

This decision will probably not manifest a single large feature, but rather become a criterion in the design of many small features (for instance, whether we choose a single-number versus a long-list display for certain evidence types).

3.4 A CONCEPTUAL PROTOTYPE OF A BADGE-BASED EVIDENCE SUBSYSTEM

Based on the three design decisions we laid out in the previous step, we now present a set of badge evidence subsystem design concepts that implement those decisions. For an illustrative example, we will follow the hypothetical case of a Loops Programming badge in an introductory computer science learning context, as it attempts to make the argument that the earner possesses the (relatively basic) programming skills needed to "Use loops to repeat sequences of commands" and "Use conditions to end a loop at the appropriate time". The purpose of this section is not to present an optimal design solution, but rather to illustrate and work through an additional layer of details. We thus frame the framework described below as a "conceptual prototype".



Loops Skill | Return to Skills List »

Use Loops to repeat program behaviors, but stop looping when appropriate.

Activity XP:

Shareable URL

http://localhost:5000/shareable

View Shared Page

Level: Bronze

View Skill Evaluation Rubric »

Requirements

Experience (XP)

level	XP
★ Bronze	100
★★ Silver	300
★★★ Gold	1000

Your score

50 XP

Beginner

Artifacts

level	pt
★ Bronze	1
★★ Silver	2
★★★ Gold	3

Your best upload

2.0 pt

[add upload]

Silver

Endorsement

You have been endorsed for this skill

3.0 pt

6/8 (75.0%)

[view] | [remove]

Gold

Exams

level	%
★ Bronze	70
★★ Silver	80
★★★ Gold	90

You have not yet attempted this exam.

Beginner

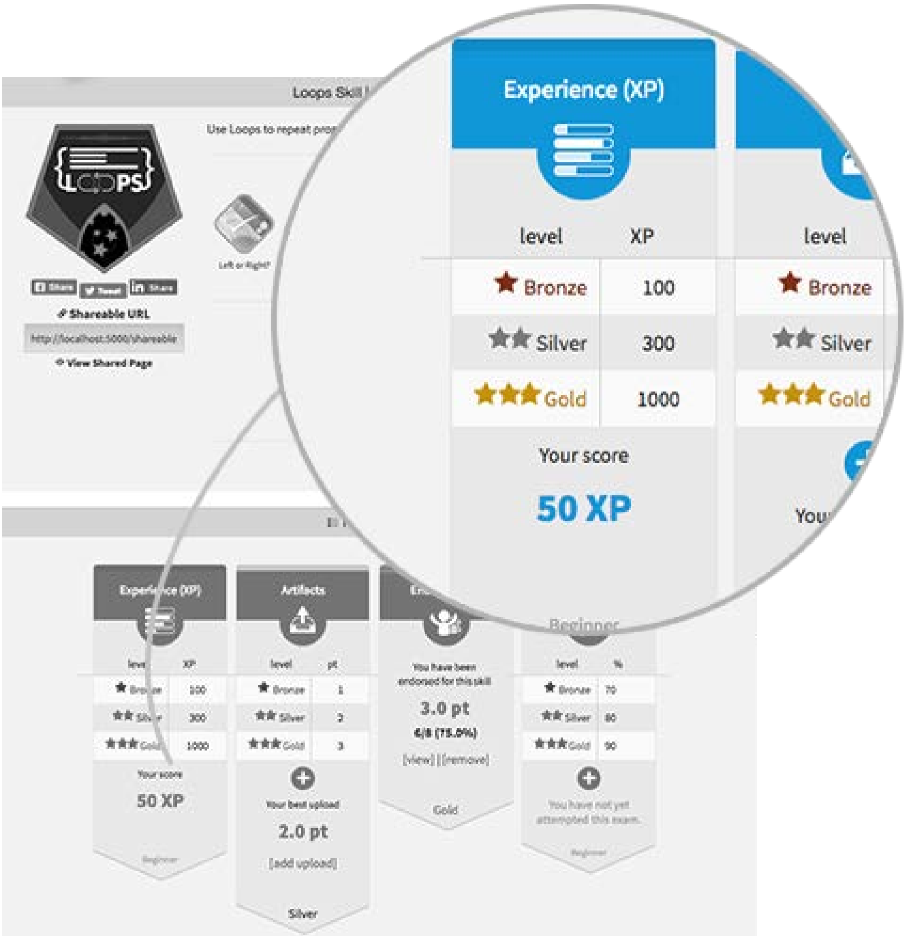
Figure 4. A composite overview of the proposed conceptual prototype.

3.4.1 Four Dimensions of Evidence

Based on Decision 1, to implement a multidimensional representation of evidence based around the types of evidence and amount of each, we have formulated a set of four major evidence categories. These categories are selected based on a combination of practical concerns, and the principle that greater epistemological diversity of evidences provides better triangulation.

3.4.1.1 XP: Experiences and experience

The XP category captures the amount of relevant experience in completing skill-relevant tasks that the learner possesses, and represents it summarily as a “XP” or eXperience Point total (drawing on a popular video gaming convention).



Experience (XP)

level	XP
★ Bronze	100
★★ Silver	300
★★★ Gold	1000

Your score

50 XP

Figure 5. An example of the XP tracker for a hypothetical learner’s Loops Programming Badge.

Our Loops Programming learner might earn 10 XP for completing an online learning module about loops, 15 XP for checking in at a hands-on programming workshop event, and another 25 XP for competing in a robotics programming competition (plus perhaps some bonus points for placing well in the competition). These would be pooled into a single summative "XP" statistic (see Figure 5).

Evidence of this type implicitly makes the claim that the badge holder has engaged with (and succeeded at) skill-relevant activities. This connects epistemologically to the main assessment claim that "the badge earner has X skill" largely through associationist theories of conditioning (training the brain to produce the right responses to task-relevant stimuli), rehearsal (repeated exposure to the correct problem-solution pair improves ability to recall the correct course of action in the future), and the motivational theory of behavioral engagement (higher levels of participation in school activities predict higher learning outcomes). These theories suggest that greater levels of experience predict greater proficiency of application, application under diverse contexts, and ability to build on these experiences to facilitate future learning. Summarily, the more one participates in (and eventually completes) skill-related activities, the better one is expected to become at them.

As a metric that lends itself to quantifiability, it is important to determine a rough scale for XP – that is, to set some common expectations around "what 100 XP means". Since we expect XP to relate to the same kinds of outcomes as rehearsal and behavioral engagement, we will use those outcomes to establish an outcome scale. Specifically, higher XP should correspond to greater ease of recall, greater future participation, and greater spontaneous recall. We therefore define the XP reference scale in terms of these quantities (albeit with initially arbitrary cutoffs that can be refined with testing):

- A learner at the 100 XP mark should be able to apply the skill with prompting and assistance

- A learner at the 300 XP mark should be able to apply the skill with minimal guidance, when prompted
- A learner at the 1,000 XP mark should be able to apply the skill fluently and spontaneously in novel situations

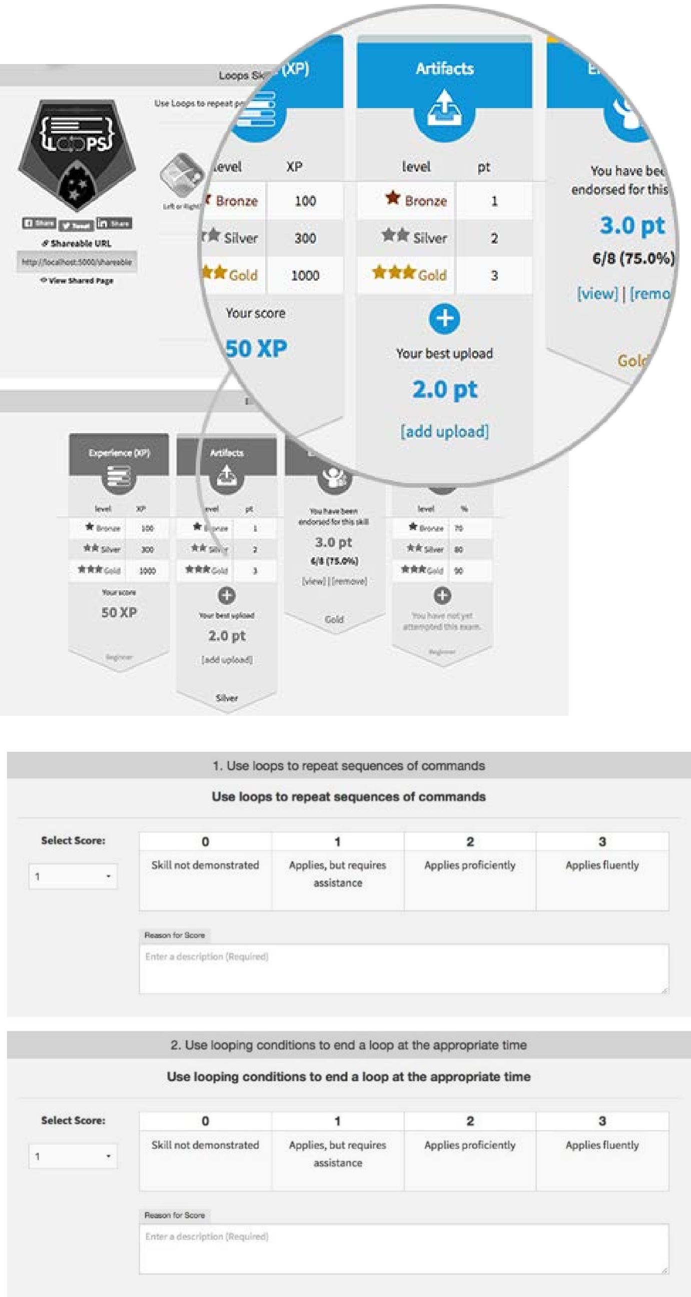
Strengths of the XP metric are its easy quantifiability (number of XP) and summarizability (as a single number). When skill-relevant activities can be identified in advance, XP tracking is easily automated. Such systems might even incorporate, e.g., diminishing point returns on highly scaffolded activities for learners who already have high levels of XP. The XP mechanic also allows the incorporation of badges from other badge systems as a form of evidence, as many settings "badge" activity completion. There are two primary weaknesses of XP as a form of evidence. The first is opacity as to exactly what kinds of activity the student engaged in, heavily favoring quantity over quality. The second is opacity of methods for completing tasks; often tasks can be completed through both use of the indicated skill and alternative brute-force methods (e.g., guess-and-check or asking for outside assistance, which learners do for many reasons; see Baker et al., 2008).

Finally, it is worth noting that interpretability of "XP" leans heavily upon the framing of the badge as a whole. Fluency in "loops programming" is easier to characterize and understand than fluency in "organizational leadership", partly because the latter is so broad. This comparison also reiterates that point values are always relative to badge scope; a task worth many points toward a narrowly-defined skillset would be worth only a few toward a broader one.

3.4.1.2 Artifacts: Contextualized work product examples

The Artifacts category captures specific concrete instances of submitted student work. Students describe and self-rate their submissions according to a badge-attached rubric upon submission. This rubric calls attention to salient features that demonstrate the skill, for both students and evaluators. It also serves as a first-order filter for quality and relevance of the submission to that skill (i.e., a student who is giving low ratings to their submission across all categories should recognize it as a poor fit for evidence toward that skill). Evaluators can subsequently interrogate the self-ratings along with the artifact as an estimate of a student's understanding of the skill itself.

A Loops Programming badge-seeker might submit an annotated copy of her source code from a class project as an Artifact, and rate herself on rubricized dimensions of "Using loops to repeat sequences of commands" and "Using conditions to end a loop at the appropriate time". Both the artifact and the rubricized rating would be made available under the "Artifacts" evidence category (see Figure 6).



Experience (XP)

level	XP
★ Bronze	100
★★ Silver	300
★★★ Gold	1000

Your score

50 XP

Beginner

Artifacts

level	pt
★ Bronze	1
★★ Silver	2
★★★ Gold	3

Your best upload

2.0 pt

[add upload]

Silver

You have been endorsed for this skill

3.0 pt

6/8 (75.0%)

[view] | [remove]

Gold

level	%
★ Bronze	70
★★ Silver	80
★★★ Gold	90

You have not yet attempted this exam.

Beginner

1. Use loops to repeat sequences of commands

Use loops to repeat sequences of commands

Select Score: 1

0	1	2	3
Skill not demonstrated	Applies, but requires assistance	Applies proficiently	Applies fluently

Reason for Score

Enter a description (Required)

2. Use looping conditions to end a loop at the appropriate time

Use looping conditions to end a loop at the appropriate time

Select Score: 1

0	1	2	3
Skill not demonstrated	Applies, but requires assistance	Applies proficiently	Applies fluently

Reason for Score

Enter a description (Required)

Figure 6. An example of the Artifact evidence interface and rubric summary display for the Loops Programming Badge.

Clicking the link gives the viewer access to the uploaded file (in this case a piece of source code), a copy of which is permanently stored on the server.

Epistemologically, evidence of this type draws on the same assessment traditions as portfolio assessment: the objects in the portfolio directly demonstrate advanced examples of skill, and a well-constructed rubric maps relevant features of the skill to relevant qualities of the objects. The rubric-rated work product allows scoring of skill-relevant performance in an authentic context.

The validity argument from this data is connected to: 1) the transparency of skills in submitted objects; 2) the rubric's relevance in representing the "right" qualities of the knowledge or skill for rating for submitted objects; and 3) the reliability and trustworthiness of the rater to give accurate ratings. The strengths of this approach lie in its dual provision of a work product with its context intact, alongside a scoring system mapped directly onto the skill claim being made. It further allows evaluators to inspect the quality of the work by examining the artifact itself. These speak directly to the earner's ability to apply and reapply the skill proficiently. The primary weaknesses of this evidence type are its vulnerability to unreliable self-raters, dependence upon the expertise of the person viewing the badge evidence (i.e., non-experts will not be able to understand the evidence), and the time-consuming nature of inspecting the artifact itself, especially for more complex artifacts (i.e., few people will take the time to investigate details for objects with a large amount of detail, such as thousands of lines of code).

3.4.1.3 Endorsement: Expert and participant verification

The Endorsement category captures assessments of intangible expertise conferred by experts and peers within a practice space. Teachers, mentors, and peers are solicited to provide a short online endorsement of the student's proficiency at the given skill, including both a rubricized evaluation and a short written statement. This replicates, to some extent, the recommendation-writing system embedded in both college and job application processes. Judgments can be based in the

Artifacts within the badging system (i.e., adding an outside evaluation of the same submitted object) or based in observation of behaviors or discussion (i.e., connected to very different sources of evidence).

The Loops Programming badge earner might solicit an Endorsement from the professional programming mentor on her robotics team. The mentor would log in to the online system, and rate the student on the rubric for the Loops Programming skill ("Using loops to repeat sequences of commands" and "Using conditions to end a loop at the appropriate time"). If the mentor has registered correctly within the system, the Endorsement is marked as being "by an expert" (see Figure 7).

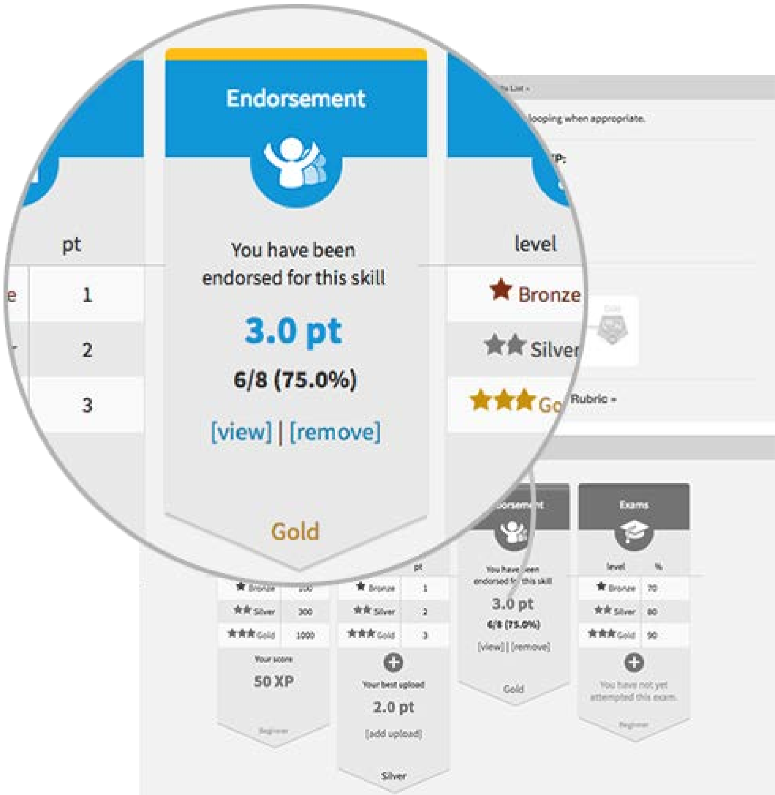


Figure 7. An example of the Endorsements this learner has received for Loops Programming. The rater was identified as an Expert based on credentials within the system.

Evidence of this type contains two features with different epistemological roots: the rubric, and social sources of endorsement. The use of the rubric maps scoring onto evaluation of relevant features of the skill, as it did in the Artifact category (the rubrics should be equivalent for this reason). The scoring process for the rubric, and the provision of the written recommendation, however, tap sociocultural mechanisms for evaluating skill.

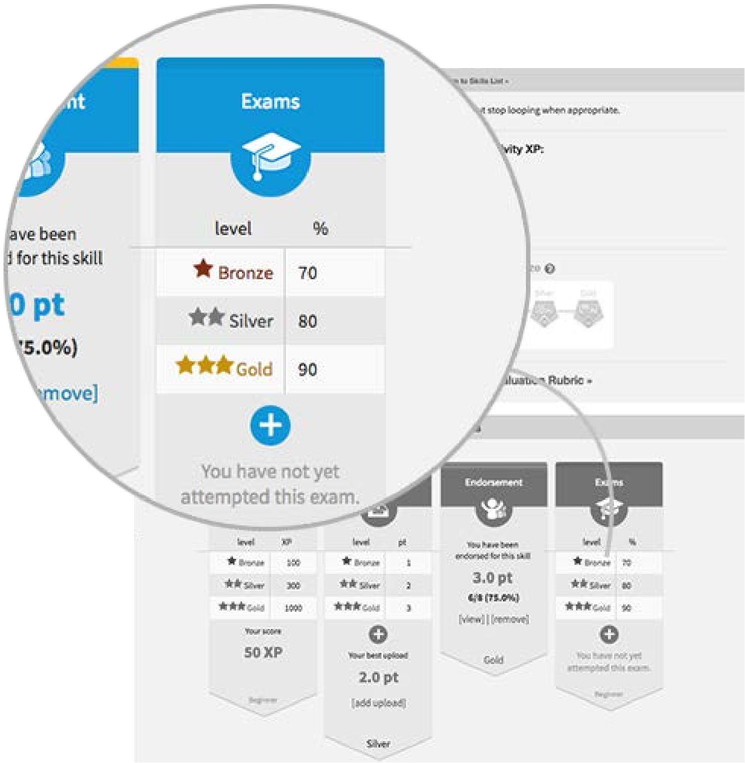
Epistemologically, there are certain kinds of skill which are notoriously hard to evaluate; tacit knowledge, for instance, refers to the unspoken and generally unmeasurable expertise that certain individuals have for making good decisions under poor information conditions (Collins, 2001). Army officers making difficult discipline and personnel decisions display a remarkable degree of similarity and expert consensus, but the underlying skill is exceedingly difficult to measure through means other than agreement among those experts (Schunn, McGregor, & Saner, 2005; Sternberg & Horvath, 1999). In sociocultural frameworks such as Situated Learning (Lave & Wenger, 1991), acceptance by the community of practice leading to increased participation *is* the mechanism by which expertise is gained.

The major strength of this evidence category is in capturing intangible, yet historically reliable, assessments of expertise by knowledgeable others within the learning space. This type of evidence is also uniquely positioned to reflect on certain types of skills which are impossible to evaluate through other means, such as collaboration quality, which is inherently social. The primary weakness of this approach is that it is difficult to establish the legitimacy of a rater, without the argument quickly becoming circular. This is particularly difficult when dealing with “non-expert” raters, e.g., peers, whose actual level of expertise (and hence reliability) might be quite low. We report this relationship (e.g., “Expert” or “Peer”) in the interface alongside the endorsement, so that an evaluator can choose to take this into account.

3.4.1.4 Exams: Transfer to strategically chosen tasks

The Exams category captures learner performance on designed measurement tasks. Implicitly, this tests students' ability to transfer their learning to the testing context. Students take a scored exam of some sort – most commonly, this will be of a traditional examination format, in which students respond to crafted prompts designed to measure one or more validated skills. In the case of exams that measure multiple skills, only the items relevant to the badged skill would count toward this score.

The Loops Programming badge might accept and electronically import exam scores from a list of trusted sources, such as vetted online courses and AP exams. These data providers would report a loops-relevant subscore, or scores of exams wholly relevant to the topic of loops programming (e.g., a Loops Chapter Exam). The Loops Programming badge interface displays the list of acceptable exams, and provides a link to at least one free online exam option, as well as the highest available score among the eligible exams (see Figure 8).



level	%
★ Bronze	70
★★ Silver	80
★★★ Gold	90

You have not yet attempted this exam.

Figure 8. An example of the Exam evidence summary for the learner’s Loops Programming Badge. If the exam has a viewable sample item/page, clicking the link the opens it for inspection.

Epistemologically, evidence of this type draws on whatever measurement techniques are “baked in” to the exams themselves. Typically, these draw upon cognitive theories of skill development and transfer (“someone who knows X will answer A for test question item I”), backed by traditional methods used to validate such items, such as Item-Response Theory. Intelligent Tutoring Systems often use evidence of this type.

The main strength of this approach is its ability to incorporate traditional measurement media as a form of evidence. They are as valid as their own methodological backing, and lend that validity to the badged evidence pool. They also allow for developing individual items that target each aspect of a large skill to allow for ‘complete’ coverage of the skill, and they encourage learners to completely master the domain rather than focusing on areas of interest.

Finally, they also carry with them the cultural and policy "weight" of these exams as they are used in the world today – insofar as the Chapter Exam is trusted as a measure of knowledge or skills today, adding it as evidence to a badge increases the badge's evidentiary warrant toward that knowledge or skill.

The main disadvantages of this evidence type are logistical: finding test items that tap a given skill and disentangling a skill-specific score from a larger exam is difficult to scale as a general practice. In many cases, standalone exams are not well-validated, capturing only superficial aspects or only declarative knowledge related to the skill (e.g., memorized answers to familiar questions), even though they may be accepted in practice.

As a final note, some exams may allow only a limited number of attempts in order to prevent users from seeking high scores by "brute force" retaking of the test. The Exams evidence construct is agnostic to this choice, leaving the decision to the individual exam administrators so as not to impinge their authenticity. While this creates the possibility that a user could be stuck with an irrevocably low score on a given test due to its strict retake policy, the user's ability to select from among multiple exams for the Exam evidence allows a workaround (albeit perhaps using a more onerous or less prestigious exam).

3.4.2 Representing "Composite Claim Quality"

Decision 2 (with addendum 2b) in the previous section task us with designing a system element that recognizes and represents a sliding scale of evidentiary warrant, and simultaneously relates strength of warrant to stakes of decisions.

Intuitively, we will want to pick a single design element to represent both *strength of evidence* and *suitability for higher-stakes decisions*. We want to pick some visual device that we can use to represent higher vs. lower levels of this shared “claim quality” dimension. However, in doing so, we should also consider other “qualities” of a badge claim that users might confuse it with. Two in particular seem to be likely sources of misinterpretation: One quality that badges frequently indicate is “more advanced skillsets”. This dimension is often represented by badge material (e.g., bronze vs. silver vs. gold), badge size, or cumulative marks such as stars. A second quality is “more advanced proficiency within the skillset” (e.g., very high levels of proficiency with beginner-level coding techniques).

The first potential confound is neatly separable. “Proficiency at more advanced versions of the skill” is better represented by a different claim, as it represents different knowledge and techniques that should be separately represented and assessed.

The other – more advanced proficiency within the skillset – maps nicely onto an evidentiary dimension we have already identified: it is the same as the “level of proficiency” that our multidimensional framework represents as rubric and XP scores. Furthermore, it correlates well with the notion of “higher stakes” – decisions such as hiring and college admissions are likely to want higher proficiency levels as well as more evidentiary certainty. Conversely, a claim of a higher level of skill deserves to be inspected more carefully and backed by stronger evidence, especially since knowledge of “all” of a skillset is likely to need a greater quantity of evidence to establish simply due to the larger “surface area” of the knowledge being claimed. Thus, we include this third dimension of “level of proficiency claimed” as a third dimension sharing the same design feature as the first two. Our composite quality to be represented is now composed jointly of “strength of evidentiary warrant”, “suitability for higher-stakes decisions”, and “proficiency level claimed”. For our design, we elect to use Badge Levels of bronze, silver, and gold to indicate the composite quality of *stronger assessment argument*, *suitability for higher-stakes decisions*, and *higher levels of proficiency being claimed*.

This is somewhat complicated by the fact that not all types of evidence may be available or suitable for all types of skill or knowledge claims. For instance, a “collaboration” skill will probably not involve an Exam. This means that such a skill has, at most, three possible types of evidence. Other skills may not have reasonably presentable work products (perhaps projects are too large, too small, or too confidential). Some may not have quantifiable “experiences” because they are innumerable, pervasive, or untrackable. Learning done by solo learners online may not have anyone to endorse them. In short, our system must be robust to a number of potential evidence types that varies anywhere from 1 to 4.

We therefore define our “composite quality” index to include the following levels, using relative counts of evidences, rather than absolute counts:

- A **Gold badge** provides **all possible evidence types**, with scores of **80% or higher** on all available rubrics and exams.
- A **Silver badge** provides **all but one of the possible evidence types**, with scores of **70% or higher** on all included rubrics and exams.
- A **Bronze badge** provides **at least one type of evidence**, with a scores of **60% or higher** if it is a rubric or exam.
- “Corner cases”: For badges which permit only two types of evidence, Silver is omitted. Badges which have only one evidence type available are referred to as “binary badges”, and they can only be earned or unearned.

3.5 RECOMMENDATIONS AND FUTURE RESEARCH

In this chapter, we have laid out a design blueprint for an assessment system rooted in the provision of evidence in and through digital badges. We began with theoretical foundations and practical use cases, and from these derived design principles, then a conceptual prototype. There remains work to be done in implementing the prototype design. For instance, the user interface design of the system plays a key role in framing and explaining the evidence requirement and submission system to both badge earners and viewers. Given the complexity and novelty of the proposed arrangement compared to conventional "one and done" exams and certification routines, completing the full interface is no mean feat. An adequately scalable technology platform would also be required to host both the evidence-collecting activities (e.g., endorsing) and the resulting badge data indefinitely.

At an organizational level, there is the challenge of identifying "expert" individuals who should appear as such when endorsing learners. Would such a qualification be imported through an existing registry of sorts (e.g., national teacher councils), or would they be "bootstrapped" into the system by some form of qualification exam? There is also a short-term need to either develop in-house exams for skills, or vet existing exams to determine which ones can qualify as Exam Evidence, followed by the need to establish a digital data pathway for importing learners' scores and associating them with the correct users in both systems. This activity extends into a long-term need to establish partnerships with commercial testing services and perhaps even individual states or school systems.

With these things in place, however, additional opportunities also open up. The digital nature of badges and of the evidences allows for high levels of integration into, e.g., intelligent tutoring or game-based systems which can automatically award XP, upload student work, and prompt teachers to provide endorsement at opportune moments in the learning process (for an expanded overview of this topic, see Ososky, 2015). Social media integration could enable additional ways of acquiring evidence; perhaps community-based “popular” endorsement by large numbers of peers could provide an alternative to “expert” endorsement for certain skills.

These are only a few of the assessment opportunities that could be tapped with digital badges. Nevertheless, future sources of evidence, however creative, will be efficacious only if these diverse and powerful sources of evidence are harnessed through a robust framework for assessment. In this chapter, we have advanced the digital badge conversation toward the development and adoption of a principled assessment framework, for it is only with this in hand that digital badges can truly unlock their potential to inspire and reward learning.

4.0 CHAPTER 4: PERCEIVED RELEVANCE OF DIGITAL BADGES PREDICTS LEARNERS' LONGITUDINAL GROWTH IN ENGAGEMENT

A badge is a marker of accomplishment, descended from marks of pilgrimage and political affiliation in the Middle Ages, military medals, and featuring prominently in modern-day Scouting movements (Ostashewski & Reid, 2015). Badges in digital format are often associated with video game achievements, but in the last few years have been re-envisioned as a potential avenue by which real-world skills might be communicated to learners, potential employers, and many other players in the world of education and credentialing. “Open” badge systems in particular – sociotechnical ecosystems without burdensome restrictions on who can issue badges – have been proposed as a catalyst for a wholesale change in the world of education (MacArthur Foundation, 2011; Mozilla, 2015).

On a perhaps less ambitious level, digital badges are also widely presumed to have some form of motivational effect. Early studies of digital badges found mixed effects of badges on both learning and motivation (Falkner & Falkner, 2014; Filsecker & Hickey, 2014). More granular analyses suggest that a three-way interaction of learner knowledge, motivation, and badge system design may underlie the effects of digital badges – that is, certain types of badges will have certain effects on certain learners (Abramovich, Schunn, & Higashi, 2013; CREATE Lab, 2015).

Qualitative studies suggest that this individual-level effect may manifest as differences in learners' perceptions of badge systems' relevance (Suhr, 2014; Davis & Singh, 2015; Wardrip, Abramovich, Bathgate, & Kim, 2016). Under this hypothesis, the degree to which individuals perceive a program's badges as worthwhile and relevant to them subsequently affects the degree to which they engage with that program. However, this factor is seldom modeled in quantitative work around digital badges.

This study builds on that work by collecting and analyzing longitudinal data from a middle school robotics programming course, in order to examine whether this effect is (i) evident over time; (ii) distinguishable from other motivational factors; (iii) has similar effects by sex, age, and race; and (iv) reciprocal, implying that positive badge perceptions may form a "positive feedback loop" with engagement that could result in increasing gains over time. We investigate these phenomena in the context of a digitally-badged computer programming module that is hosted online.

4.1 BACKGROUND

4.1.1 Digital Badges

At a mechanical level, a digital badge is remarkably simple. In response to a certain event, a computer program generates a packet of data containing an image, a few text fields, and possibly some metadata linking to other data (Hamari & Eranti, 2011; McDaniel & Fanfarelli, 2016; Open Badges Project, 2017). The rest is semantics: digital badges are meant to symbolize achievements, the latest development in a long and storied history of symbols of accomplishment in human culture (Ellis, Nunn, & Avella, 2016; Halavais, 2012; Ostashewski & Reid, 2015). In the past few years, attention has turned particularly to the use of digital badging in the world of education. Such systems draw inspiration from a number of sources – Scouting traditions like the Boy Scouts and Girl Scouts, online reputation systems like the one used to indicate high-quality posters on the community Question & Answer site Stack Overflow (https://www.stackoverflow.com), and video game achievement systems like Xbox Achievements (Jakobsson, 2011; Jakobsson & Sotamaa, 2011).

Thus, badges' bits and bytes are ascribed powerful meaning: one text field describes an achievement, another gives it a name; the image is a visual representation of that accomplishment. Criteria for earning the badge are specified. Metadata is attached as evidence that recipients have met the criteria. Names of issuing parties and lists of awardees are recorded. Badge earners take copies of these digital records with them, and a free-standing credential is born. Deliberately absent from this process are any restrictions on who badge issuers might be (Open Badges Project, 2017).

This combination of technical simplicity, semantic depth, and mass availability has been argued to position Open Badges as a disruptive innovation (MacArthur Foundation, 2011; Mozilla, 2015). By "unbundling" the traditional package of curriculum, instruction, assessment, and credentialing, digital badges challenge the de facto monopoly on educational offerings by institutions of formal education, a goal referred to as the *democratization* of learning (Grant, 2016, p.8). In their place, learners will construct portfolios – or "backpacks" (Mozilla, 2015) – of badges, acquired on self-discovered or suggested pathways through learning opportunities in diverse settings.

A related goal – or perhaps a side effect of fragmentation in a many-provider world – is that the "grain size" of credential claims become smaller, moving away from diploma-scale claims about general skillsets in favor of *microcredentials* indicating proficiency at specific skills, tools, or technologies. This finer granularity in turn allows more specific evidence of proficiency to be presented via electronic portfolios, which employers overwhelmingly prefer to transcripts (Hart Research Associates, 2015). Badges can also represent skillsets which diplomas typically do not, including the "soft skills" like teamwork and problem solving often sought by employers (Barton, 2006; Heckman & Kautz, 2012; Whitmore & Fry, 1974). A small number of K-12 school systems have already begun adopting badges toward this end, documenting 21st Century Skills in collaboration with local employers (Derryberry, Everhart, & Knight, 2016).

Finally, central to the current study, digital badges have been largely expected to have effects on learning and motivation. The educational badges concept gained popularity as part of a larger wave of *gamification* research seeking to bring techniques used in video game design – seen as successful in motivating students toward long-term engagement – to the design of educational systems (Deterding, Sicart, Nacke, O'Hara, & Dixon, 2011). This approach has seen very limited success overall, but there is some reason to believe that badges may have unique merit, effecting change through motivational mechanisms such as identity (Gibson, Ostashewski, Flintoff, Grant, & Knight, 2015), interest (Eccles, 2009), proximal goal-setting (Antin & Churchill, 2011; Rughinis, 2013), and value of various types (Wigfield & Eccles, 2000).

4.1.2 Theoretical Framing

A great deal of badging research has attempted to measure digital badges' impact wholesale on learning, motivation, and behavior, either as a standalone intervention or as part of a broader approach of gamification (e.g., Filsecker & Hickey, 2014; Hanus & Fox, 2015; Hew, Huang, Chu, & Chiu, 2016). This approach has generally produced mixed findings (Falkner & Falkner, 2014). Subsequent studies have made greater headway by changing the central framing of the question from *whether* badges work, to *when and where* they work (Itow & Hickey, 2016), and *what kind for whom* (Abramovich, Schunn, & Higashi, 2013; Reid, Paster, & Abramovich, 2015). This shift toward a more situative perspective – one that accounts for the contextual factors around badges' implementation, including those of the individual learner – produces a clearer picture of factors for designers to take into account, and for researchers to measure.

4.1.2.1 Perceived Badge Relevance

The current study explores one underexamined element of this system: individuals' perceptions of badges. This topic is implicit in nearly all badge studies, yet it is curiously understudied, particularly in quantitative work. The relevance of participant perceptions is somewhat better established in qualitative work: Suhr (2014) noted some members of an online community simply dismissed its badges as undesirable because they were earned for achievements they considered trivial; Davis and Singh (2015) report on a program whose badges were seen as pointless, ignored in favor of grades, and subsequently had no effect on students' activity. Wardrip, Abramovich, Bathgate, and Kim's (2016) interviews connect participants' perceptions of non-triviality to meaning and value. We hypothesize that these factors – finding the badges desirable, valuable, and contextually meaningful – should predict the degree to which the badges affect their behavior in the program overall. We refer to the overall construct as "Perceived Badge Relevance" (PBR): the degree to which an individual believes a program's badges are relevant to them.

Yet while quantitative studies have frequently investigated badge-level factors and outcomes of badge use, the critical *perception* link between participants and badges has been operationalized by only a handful of studies (e.g., Reid, Paster, & Abramovich, 2015). And even these did not examine the entire chain to connect learners, their beliefs about badges, and subsequent behavioral outcomes.

PBR as a construct is inherently synthetic in nature – there are a great many reasons why an individual might find a badge relevant to them or not – yet we believe that a theoretical construct of "perceived relevance" may usefully abstract their common element of personal connection to the badges in a way that may increase engagement in learning activities.

4.1.2.2 PBR and engagement

Engagement is a term that has been used to refer to many different constructs. Here we focus on a form that is theoretically and practically distinct from motivation – in fact, simply put, engagement as the *consequence* of motivation. We use a formulation of engagement in educational contexts that originated with Fredricks' (2004) conceptualization of "school engagement" as encapsulating a student's involvement and participation with school. One of the primary features of this model is that engagement is a coherent whole, but composed of three qualitatively different *types* of engagement. *Behavioral* engagement includes learners' choices of activities, and actions taken within them. *Cognitive* engagement is learners' mental engagement with problems and tasks. *Affective* engagement includes feelings and attitudes that learners have toward the task.

For the current research, we adapted a specific conceptualization of Engagement developed by the Science Activation Lab to be appropriate to both out-of-school programs and mixed contexts (e.g., Dorph, Cannady, & Schunn, 2016; Ben-Eliyahu, Moore, Dorph, & Schunn, 2018). This conceptualization focuses on engagement with activities (rather than the broader context of school or school subjects), and characterizes engagement in the moment (rather than as a long-term quality). This formulation is appropriate to a wide variety of digital badging contexts in that they will all have activities that must be completed in order to obtain the badges.

We hypothesize that students who find a program's badges more personally relevant (i.e., have higher PBR) will be more willing and able to engage in the program's activities. Specifically, higher PBR should indicate an increase in receptiveness to overall effects of badges. For example, if we suppose that a badge supports goal-setting, and thus self-regulation (Charleer, Klerkx, Odriozola, & Duval, 2013; Pintrich & De Groot, 1990), then students who perceive a set of badges to be more relevant to them are more likely to adopt self-regulating behaviors, which subsequently lead to increased behavioral and cognitive engagement. A second example involves the direct seeking of badges. Learners who find the badges personally relevant are more likely to attempt to earn them. By nature, earning badges means completing domain-relevant tasks, so learners acquiring the badges will be at least behaviorally engaged in domain activities. Learners who find the badges relevant may be more attentive to activities in the domain, making them less likely to be bored or inattentive. Thus, we propose that higher PBR increases a learner's receptiveness to badges' effects; because badges are generally designed to increase learner engagement, this results in increased engagement with domain activities. However, note that theoretically, some have argued that badges can be thought of as kind of extrinsic reward (Resnick, 2012), and some extrinsic rewards have been associated with reductions in engagement and learning (Ryan, Mims, & Koestner, 1983; Deci, Koestner, & Ryan, 2001).

Higashi and Schunn (in review) found in a multi-level analysis that individual learners' PBR predicted their engagement across a wide variety of programs. The effect was positive and consistent across 45 summer programs, even when controlling for age, program size, and perceived success in those programs. However, the dataset used in that study was cross-sectional, and inherently unable to distinguish whether PBR is actually associated with change processes, or if the two are simply correlated at single moments in time.

4.1.2.3 PBR and motivation

We argue that PBR is conceptually distinct from domain motivation because it is specifically *about badges*. But there is an important empirical question in studying the relationship of PBR to engagement that involves motivation: can the relationship between PBR and engagement be fully accounted for by other motivational factors? That is, are perceptions of badge relevance simply proxies for other, well known motivational factors such as interest, and interest is what drives engagement? Higashi and Schunn found that perceived success during program activities, another motivational construct, was related to PBR, but each predicted distinct variance in engagement. However, their study did not include persistent long-term motivational covariates such as interest or identity, each of which is a known predictor of engagement, and likely correlated with PBR.

Domain Identity. As social signifiers, Badges are intrinsically tied to identity. Gibson et al. (2015) suggest that badges may act through channels of personal and social identity to "assist users in building and formalizing identity in social media networks". The converse may also be true, that individuals who already see themselves as having an identity in a domain will be more inclined to display this fact to others. Thus, domain identity – in the current study, identity as a programmer – may be an important predictor of an individual's perceptions of value and desirability of badges (and hence their PBR) in that domain.

It will therefore be important to account for the predictive relationship between identity and engagement directly, in order to isolate the effects of PBR. Learners who engage productively in activities and settings associated with a particular discipline are more likely to continue to participate, and ultimately consider careers in those disciplines (Collins, 2006; Lave & Wenger, 1991; Engle, 2006; Aschbacher, Li, & Roth, 2010). Learners also choose to engage with activities that are compatible with their senses of social identity, which are often gendered (Kessels, Heyder, Latsch, & Hannover, 2014). Thus, we expect that individuals' level of identification with the domain of learning may predispose them toward continued increases in engagement. Note that participatory notions of identity, social stereotypes, and the "social signifier" hypothesis of badge action all include the assessments of others in the domain space (e.g., peers and teachers) as well as the individual's own.

Interest. Wigfield & Eccles's (2000) Expectancy-Value Model of motivation explains student decisions using a simple logic: students only attempt a course of action if they think it will be worthwhile in that a successful outcome would be valuable, the attempt is likely to succeed, and the relative cost of the attempt is not too high. The intrinsic (also called "interest") value component of subjective value includes the learner's interest in the learning domain in which the badges are positioned. We are particularly concerned with sustained *individual interest* (Hidi & Renninger, 2006; Schiefele, 2009) that persists over time and is reapplied across situations, rather than the more momentary and fleeting *situational interest*. Individual interest in the program domain is likely to be a stable predictor of the degree to which an individual decides to engage in program activities, but could also directly increase an individual's perception of earning domain badges as desirable. The constructs of domain interest and identity are related. An individual’s interest in a domain may be informed by his or her personal or collective senses of identity (Eccles, 2009). This raises the possibility that the two may be highly correlated in a particular context.

4.1.2.4 Badges and equity

There are many well-noted, persistent discrepancies in educational attainment and performance outcomes in the United States between groups that are theoretically entitled to equal treatment under the law, particularly across lines of race and gender. There is also pervasive demand to improve workforce development capacity by improving education for traditionally underserved populations (Committee on Underrepresented Groups and the Expansion of the Science and Engineering Workforce Pipeline, 2011; Allen-Ramdial & Campbell, 2014). In the context of computer science and STEM education, from which our data is collected, women are greatly underrepresented, as are members of minoritized racial groups. Minoritization is not the same as being a member of a statistical minority – while the latter is a simple property of population numbers, minoritization refers to the effect of sociocultural forces in marginalizing members of certain racial groups *because* of their underrepresentation (Bishop, Berryman, Wearmouth, & Peter, 2012). The exact constitution of minoritized groups thus varies by domain; in STEM fields, males and individuals of White or Asian racial background tend to be over-represented relative to the general population, and so females and individuals with non-White/Asian backgrounds would be minoritized (Burke & Mattis, 2007). If badges are to be used in real-world educational contexts, their effects must be *at least* neutral with regard to race, sex, and age.

4.1.2.5 Positive feedback loops

Researchers have theorized that recursive feedback mechanisms may central to producing the large observed differences in, e.g., drop-out rates among minority students (Cohen et al., 2009). In a recursive mechanism, small differences in a learner's initial perceptions of his or her own ability within a domain causes differences in performance on a task in that domain; the learner observes this discrepancy as one that confirms and strengthens his or her perception of his or her own (un)suitability for the environment, which leads to even larger differences in performance the next time, and so on, ultimately leading to large differences in levels of performance and (when negative) higher rates of dropping out. Investigators in the Cohen et al. (2009) study implemented a very brief psychosocial self-affirmation intervention, which disrupted a negative feedback cycle among minority college students, nullifying a large portion of the achievement gap among treated students.

We theorize that digital badges may also be a simple intervention that eventually produces large effects by producing feedback loops. However, rather than disrupting a negative feedback loop, we believe that badges may instead promote a positive feedback loop. Specifically, learners who perceive the badges as personally relevant may engage more with learning content, causing them to earn more badges and thus see themselves as belonging more in the space, in turn increasing their perception of the badges as relevant to them. Thus, the presence of badges may induce a positive feedback cycle between PBR and Engagement. If this is occurring, we should observe both predictive effects in a positive direction over time: higher PBR predicting increases in engagement, and higher engagement predicting increases in PBR.

4.1.2.6 Research questions

We have proposed that an individual-level factor we call Perceived Badge Relevance is of theoretical and practical importance in understanding the impact of digital badges in education. Past studies have been cross-sectional, and inherently unable to distinguish whether PBR is actually associated with change processes, or if the two are simply correlated at single instants in time.

In this study, we seek to answer three main research questions:

RQ1. *Does PBR predict engagement?* This breaks into two related sub-questions. *(RQ1a) Is there evidence of a longitudinal effect of PBR on engagement?* If PBR is indeed responsible for moderating the effects of badges, then these effects should be visible over time. While longitudinality is not sufficient to show causation, it is necessary; we must address the concern that the observed effect of PBR in cross-sectional data sets is simply an artifact of covarying endogenous selection. *(RQ1b) Is PBR empirically distinguishable from domain motivation and demographic factors?* Two major sources of endogenous variation relevant to the relationship between PBR and engagement are domain motivation and population bias. It is possible that perceptions of badge relevance are wholly determined by factors such as interest and domain identity, or that they simply reflect more general patterns of bias due to age, sex, or race. Additionally, badges are embedded in the programs and domains in which they are used. It is possible that learners will simply treat them as a part of those domains, and that will be no distinguishable badge-specific impact on engagement.

RQ2. *Are the effects of badges equitable?* As interventions in educational ecosystems, badges are obligated to attend to questions of fairness. We should be careful that digital badges do not widen existing inequitable achievement gaps. If certain groups are inclined or disinclined toward badges, this could become relevant.

RQ3. *Is there evidence of reciprocal effects between engagement and PBR?* Psychosocial interventions such as badges can sometimes create feedback loops that result in large changes over time, in response to relatively small and short interventions. Such a situation could hypothetically occur with badges, but to date, this possibility has been unexplored.

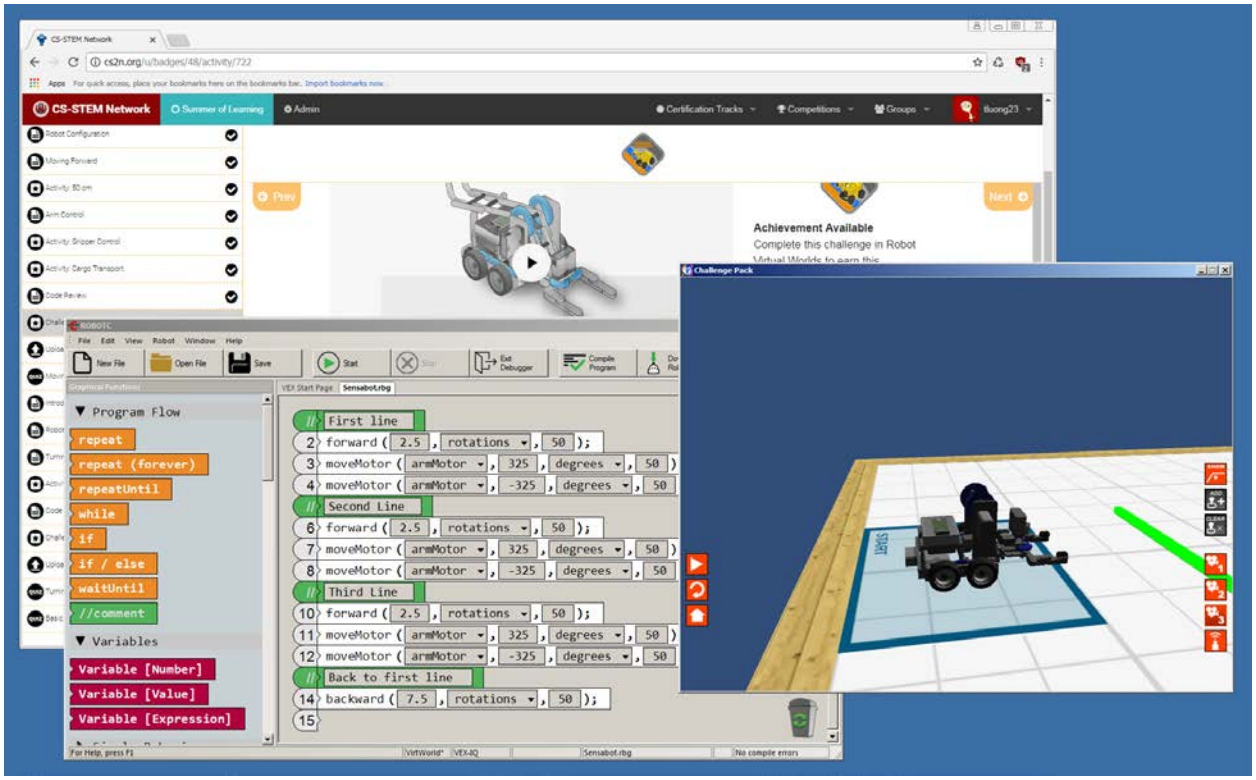
4.2 METHODS

4.2.1 Context

4.2.1.1 Online curriculum

We conducted our study of badges in the context of an online computer programming course module designed for middle and high school students. In this module, students program simulated 3D agents ("virtual robots") to complete themed tasks resembling those that real-world robots perform, such as navigation, object sorting, and disaster relief. Its primary learning objectives are robotics-related – motors and sensors – and coding-related, including command sequences, loops, and conditional statements (if-else).

Students wrote code in the ROBOTC for VEX software's graphical programming mode, which uses block-based commands similar to other popular languages such as Scratch. This software connected to an instance of a Robot Virtual Worlds simulator running locally on their Windows PC. The specifics of the robot (e.g., names of its outputs) were matched to the real-world VEX IQ robot kit commonly used in K-12 education. Video lessons in a browser provided guidance and direct instruction. Figure 9 shows the simulator interface, the coding interface, and the curriculum interface as seen by the user.



```
// First line
forward ( 2.5 , rotations , 50 );
moveMotor ( armMotor , 325 , degrees , 50 )
moveMotor ( armMotor , -325 , degrees , 50
// Second Line
forward ( 2.5 , rotations , 50 );
moveMotor ( armMotor , 325 , degrees , 50 )
moveMotor ( armMotor , -325 , degrees , 50
// Third Line
forward ( 2.5 , rotations , 50 );
moveMotor ( armMotor , 325 , degrees , 50 )
moveMotor ( armMotor , -325 , degrees , 50
// Back to first line
backward ( 7.5 , rotations , 50 );
```

Figure 9. Software interface from the learner's perspective.

Top: CS2N course materials; left bottom: ROBOTC programming software; right bottom: Robot Virtual Worlds 3D simulations environment.

Simulation activities are organized in a linear sequence of video lessons that provide instruction on a new skill, mini-challenges in which learners practice the skill, and a culminating chapter challenge which requires a thoughtful application of the skill. Lesson flow consists of a roughly repeating pattern of 1-4 interwoven videos and mini-challenges, followed by a chapter challenge. Quizzes, exams, and upload links to turn in source code are interspersed among the lesson pages. End-to-end, the course module consists of approximately 50 lessons, organized into 7 chapters, and takes between six weeks and a semester, depending on course frequency and pacing. Additional description of the curriculum can be found in Witherspoon, Schunn, Higashi, and Shoop (2018).

CS2N. Course material was delivered through an online learning management system (LMS) called CS2N: The Computer Science STEM Network. Most of its content is open to the public, and there are mix of independent learners and students working in organized (formal and informal) education settings. In formal education settings, teachers create CS2N Groups to organize their students and provision the appropriate content for them; teachers are also able to view students' progress via the badges they have earned. While all content is available to learners who are enrolled in the class, teachers regularly select only specific chapters for use in their classrooms.

4.2.1.2 Badges

In the curriculum module, students could earn badges of different types based on completing simulation activities and quizzes. Completing a mini-challenge or challenge required students to program the simulated robot to move to a certain area or transport other in-game objects to certain positions. Upon completing each mini-challenge or challenge, users were awarded a badge on-screen, and the simulation software automatically communicated this achievement to the online badging platform.

Lessons combined video instruction and mini-challenge practice opportunities on single pages by topic (e.g., "Turning in place"). A page was considered complete when the corresponding mini-challenges and challenges were completed in the simulation – this corresponded with all the associated badges being earned by the user. Each page displayed a list of relevant badges at the top, along with their completion status so that both progress and expectation were visible to the user. A sequential listing of badges and their earned status were viewable by students (self only) and teachers (all students in the class).

CS2N badges. CS2N's digital badging system is based on an underlying theory of evidence that relies on the collection of multiple, qualitatively different types of data to make the joint case that the badge earner possesses the claimed skill or knowledge: a set of relevant successful experiences (Experience or "XP"), a set of work product artifacts, and a written examination (Higashi, Schunn, Nguyen, & Ososky, 2017).

Based on testing, however, (a) it was not practical to require users to frequently upload evidence, (b) only a certain amount of collection could be automated, and (c) users did not consistently attend to differences between listed skills. Therefore, the system focused on many fully automated badges (one per completed simulator scenario) that led toward larger badges at the chapter level. Completion of the chapter level badge was contingent upon successful completion of the individual mini-challenges (XP), upload of source code for key challenges (artifacts), and attaining a passing score on a chapter exam. Finally, to reduce the informational load on users, naming conventions for concepts, content organizers, and badges were unified – for example, the second chapter is called "Robot Movement", covered commands to make the robot move, and was represented by a badge called "Robot Movement".

4.2.2 Participants

Our data was collected from 2,410 students who used the programming course module while associated with a CS2N group. Participants were located across the US, and have historically come from middle school technology, robotics, and computer science classes. A subsample of 458 students for whom we have demographic information available suggests that the sample is drawn from the general middle school population – 49% female, with a mean age of 12.6 years (SD=1.0) – and were commonly organized into classroom-sized units with a median size of 19 students (mean = 29 students), although the range of group sizes in the full sample (min = 1, max = 125) suggests that our sample also included some small afterschool programs, home users, and multi-section groupings. 12% of subsample respondents reported racial backgrounds that are underrepresented in technology fields (non-White or Asian).

4.2.3 Measures

4.2.3.1 Engagement

Engagement was measured using three Likert-scale items selected from the Activation Lab Engagement scale (Science Learning Activation Lab, 2016a). This small number was necessary to fit the data collection instrument limitation of three items per survey (see *Instruments* below). We selected the three items with the highest factor loadings on the shared "unidimensional engagement" factor in the full scale's bi-factor model. The items were (R indicates the item is reverse coded): "During this activity, I felt bored" (R); "During this activity, I felt happy"; and "During this activity, I was daydreaming a lot" (R). Response choices were 1="NO!" through 4="YES!", a Likert scale format found to have low cognitive load in younger learners, generally produce equal distance item separation in IRT analyses, and support appropriate use of means across items (Bathgate & Schunn, 2017).

In available validation data of the full instrument, Item 1 loaded very highly on the common factor (.80) and slightly (.27) on the Affective sub-dimension, Item 2 loaded highly on both the common (.62) and Affective (.64) factors, and Item 4 loaded highly on the common factor (.63) and weakly on the Cognitive-Behavioral factor (.38). Our three-item subset thus primarily represents the unidimensional Engagement construct, but includes unique variance from all Engagement sub-dimensions (e.g., the three items were not all Affective-loading items in the bi-factor model).

Because of this dual nature, the construct displayed moderate reliability according to Cronbach's Alpha in our data set (α = .72). A confirmatory factor analysis reached a similar conclusion. As we did not wish to lose the unique variance attributable to the affective vs. cognitive-behavioral distinctions, we collapsed the three items to a mean scale score, which we use to produce Engagement scores. We argue that this mean score model is preferable to modeling Engagement as only the shared variance of the three items, because doing so would largely "partial out" the sub-dimensional variance that is theoretically well-established in the Engagement construct.

4.2.3.2 Perceived Badge Relevance (PBR)

Perceived Badge Relevance is modeled as a latent factor representing the degree to which learners believed the programs' badges to be relevant to them, in terms of contextual meaningfulness, value, and desirability; these are aspects underlying badge relevance that were highlighted in prior interview studies with badge earners (Davis & Singh, 2015; Wardrip, Abramovich, Bathgate, & Kim, 2016; Suhr, 2014). PBR was measured using three items on the same 4-point [NO!-no-yes-YES!] Likert scale as Engagement: "The badges in this program make sense to me", "The badges in this program are valuable", and "I want to earn the badges offered in this program." The 3-item scale for Perceived Badge Relevance displayed high reliability (α = .85). This is similar to the fit observed in a previous study using this measure (Higashi & Schunn, in review).

4.2.3.3 Interest

To measure a learner’s interest in computer programming (the subject domain of the curriculum), we adopt a previously developed measure (Witherspoon et al., 2018), which has four Likert-scale items. It is based on the Fascination construct and measures used by the Science Learning Activation Lab (2016b) and is analogous to individual interest (Hidi & Renninger, 2006) or intrinsic/interest value (Wigfield & Eccles, 2000). The items are as follows (R indicates the item is reverse coded): “I wonder about how computer programs work.” [Never-Once a month-Once a week-Every day]; “In general, when I work on programming, I:” [Hate it-Don’t like it-Like it-Love it]; “In general, I find programming:” [Very boring-Boring-Interesting-Very interesting]; “After a really interesting programming activity is over, I look for more information about it.” [NO!-no-yes-YES!]. In our data, the scale displayed high reliability, Cronbach’s Alpha = .87.

4.2.3.4 Identity as a programmer

To operationalize identity in our lesson context of computer programming, we use the identity measure from Witherspoon et al. (2018), which is formed from four Likert Scale items: “I am a ‘computer programming person’.” [Not me—Exactly me]; and three items of the form “My [family]/[friends]/[teachers/instructors] think(s) of me as a ‘computer programming person’” [NO!-no-yes-YES!]. In our data, the scale displayed high reliability, Cronbach’s Alpha = .90.

4.2.3.5 Demographic data

Participants who completed the pre-survey (see *Instruments* below) reported their age (in years), sex (Boy or Girl), and race (White, Black or African-American, Asian, Indian or Middle Eastern, Native American/Pacific Islander, Hispanic/Latino, I Don’t Know, or Other). Multiple selections were allowed for the race response. Race responses were then processed into a dichotomous “minoritized” status factor: Not Minoritized (0) if the user selected either White or Asian among their choices, Minoritized (1) if they did not select either, and the data was treated as missing if the respondent selected “I don’t know” (even if they also selected other options).

4.2.3.6 Groups

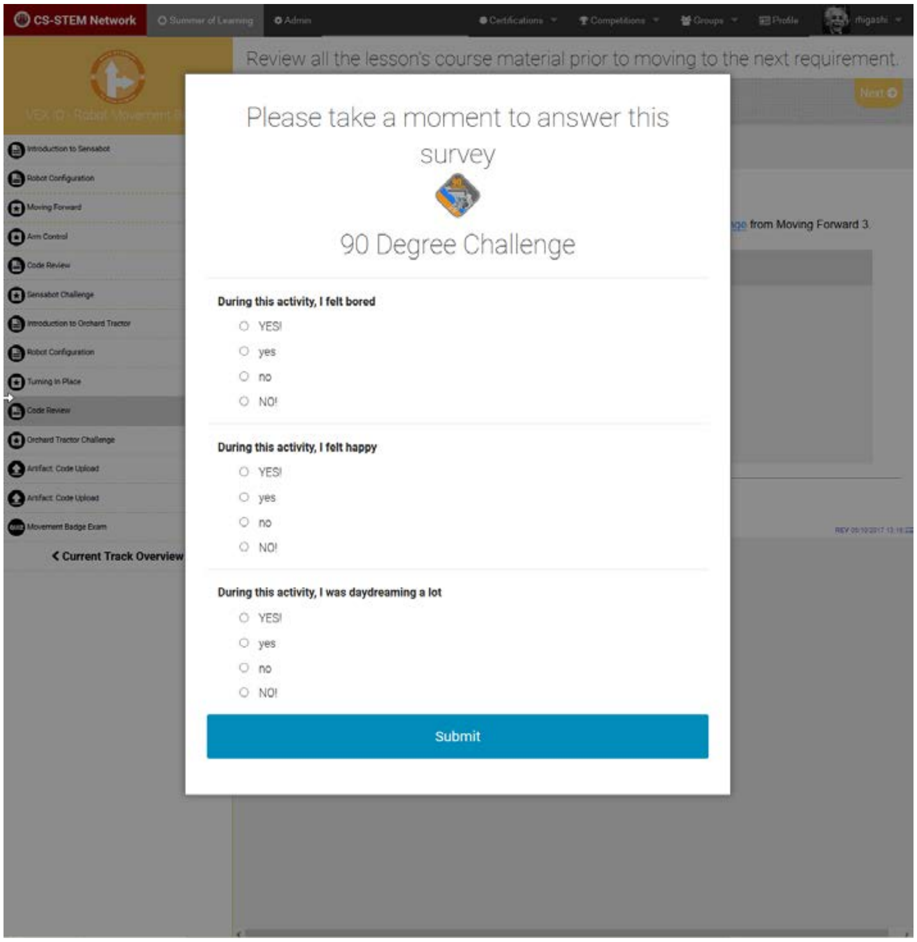
As a unit of clustering, we use the built-in Groups feature on CS2N. In general, teachers use Groups to associate users in a single class or class period (e.g., Period 8 Robotics). Thus, a teacher typically has 3-4 Groups, with each group representing an organic clustering of students, frequently by grade, subject, and/or ability level.

4.2.4 Procedure

The data for this study was acquired by two mechanisms: a three-item survey administered automatically through the CS2N platform upon completion of each lesson, and a paper pre-survey given to a subset of participants. Data were matched using CS2N account names, which students were asked to write down on the pre-surveys.

4.2.4.1 Engagement and PBR surveys

Using CS2N’s Survey feature, we measured Engagement and PBR following completion of mini-challenges and chapter challenges. Surveys display the relevant just-earned badge at the top to establish context, as shown in Figure 2. To insure high response rates and meaningful responses, each survey was limited to 3 items. Therefore, the two surveys were administered in roughly alternating order: one survey containing the three Engagement items, and one containing the PBR items. More specifically, the Engagement survey was administered to students immediately after the completion of each Mini-Challenge (i.e., when Mini-Challenge badges are earned), and the PBR survey was administered after each Challenge is completed (and accompanying badge issued). This design was particularly useful testing temporally ordered predictions of one construct to the other (i.e., PBR predicted future engagement and engagement predicting future PBR).



Review all the lesson's course material prior to moving to the next requirement.

...ge from Moving Forward 3.

Please take a moment to answer this survey

90 Degree Challenge

During this activity, I felt bored

○ YES!
○ yes
○ no
○ NO!

During this activity, I felt happy

○ YES!
○ yes
○ no
○ NO!

During this activity, I was daydreaming a lot

○ YES!
○ yes
○ no
○ NO!

Submit

Figure 10. An Engagement survey as seen by student users.

“90 Degree Challenge” is the name of both the Badge and the mini-challenge activity. The survey is overlaid onto the first CS2N page students view after completing the activity.

4.2.4.2 Qualtrics pre-survey

10 of the implementing teachers were recruited through a professional development course volunteered to administer an additional online survey to their students prior to the beginning of their robotics unit. The survey included the Interest and Programmer Identity scales, as well as demographic information such as age, sex, and race. The demographic questions came after the motivational questions, so that the motivational items would not be strongly biased by the activation of potential gender or racial stereotypes. These surveys were matched to the CS2N Survey data using CS2N user IDs inputted by the respondent.

4.2.5 Analytic Plan

4.2.5.1 Tris

Due to the pattern of construct availability in our dataset (i.e., that mini-challenges have only Engagement data, while chapter challenges have only PBR data), our longitudinal model differs somewhat in form from a conventional growth curve model, in which all data is present at all points. We instead exploit the data's patterned availability to isolate data triads in which two adjacent data points are available for the dependent variable, along with one reading of the key independent variable taken in between them. We refer to these triangle-shaped data units as "Tris". A Tri of two adjacent Engagement measurements with an intervening PBR measurement is depicted graphically in Figure 11. As it comprises a temporal sequence of Engagement-Badges-Engagement measures, we refer to it as an EBE Tri.

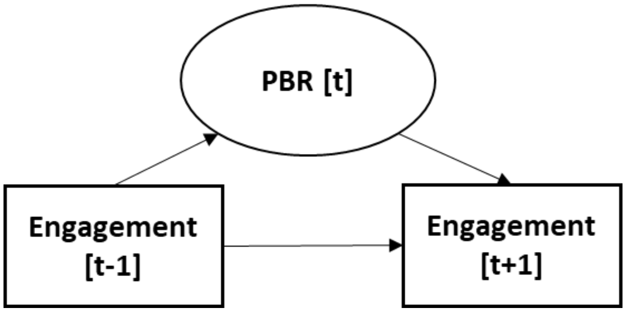


Figure 11. An Engagement-Badges-Engagement (EBE) Tri.

Tris are the fundamental unit of analysis for our models. As the formulation suggests, the Tri is designed to allow longitudinal inference about the relationship of an independent variable at some point in time (t) to a dependent variable's value at a subsequent point in time (t+1), while controlling for an individual's prior value on the dependent value (at time t-1). An individual who completes multiple consecutive activities would generate multiple Tris, nested within that individual.

4.2.5.2 Tris in the CS2N Curriculum

Importantly, all Tris are considered to be comparable to each other. One threat to this assumption is that Engagement or PBR data from adjacent data points might not be comparable due to innate, idiosyncratic differences in content. An inspection of mean levels of key variables (Engagement and Perceived Badge Relevance) suggests that this is not an issue in our data – mean engagement and mean PBR did not vary substantively between lessons. This lack of volatility suggests that no radical difference in interpretation was taking place, and that there were no large differences in, e.g., quality or boringness between lessons (see Appendix A). Nevertheless, we test this assumption whenever possible in our analysis by replicating our results on the subset of data from only a specific chapter Tri.

The Tri format also addresses a second set of difficulties in modeling longitudinal effects in our data: the heterogeneity of curriculum usage stemming from CS2N's flexibility. Whereas a typical longitudinal study might examine patterned changes in a dependent variable measured at fixed increments of time (e.g., a growth curve model), our data points are measured at fixed points of student curricular progress – and these may occur hours or weeks apart, depending on the instructional format used in that student's class. This requires us to adopt a longitudinal unit that is not *days of class time*, but instead, *lessons of curricular progress*. This notion of longitudinality is innate in the TRI format.

A third analytic difficulty arises from the fact that individual instructors on the CS2N platform have great discretion in deciding which curriculum modules to use with their classes. While a traditional curriculum efficacy study might look at gains occurring along a consistent set of lessons, instructors on CS2N routinely used only selected chapters of the material (for instance, the unit on sensors) with their classes. This form of heterogeneity is typical instructional behavior within classrooms, and therefore unavoidable without a cost to external validity. Nevertheless, selective usage induces patterned missingness in our data, which cannot be mechanically accounted for by missing data techniques such as multiple imputation or full information maximum likelihood estimation. Instead, we account for this effect by making the assumption that most curricular-assignment heterogeneity is a consequence of instructor choices on a per-classroom basis, and we include classroom-level nesting effects in our models.

Processing Tris. Because the chosen curriculum features variable numbers of mini-challenges between chapter challenges, and because we are interested in overall patterns of engagement, we collapse item scores across consecutive sequences of mini-challenges to the sequence means. Adjacent chapter challenge surveys were not collapsed, as they were typically farther apart in both content and timing; when multiple challenges occurred in sequence without mini-challenges between them, there were simply no Tris formed. We narrowed our lesson selection to include only portions which were commonly used during the data collection period, reducing the number of tracked lessons from 7 to 5. One bug occurred in data collection occurred, in which a challenge at the end of the Sensors section triggered the Engagement survey rather than the PBR survey – we simply treated it as an Engagement reading. The resulting data is composed of the expected alternating sequence of engagement and PBR measurements, yielding four potential EBE Tris (and five potential BEB Tris of relevance to RQ3) per learner over the length of the curriculum module. A map of the effective data collection pattern can be seen in Figure 12.

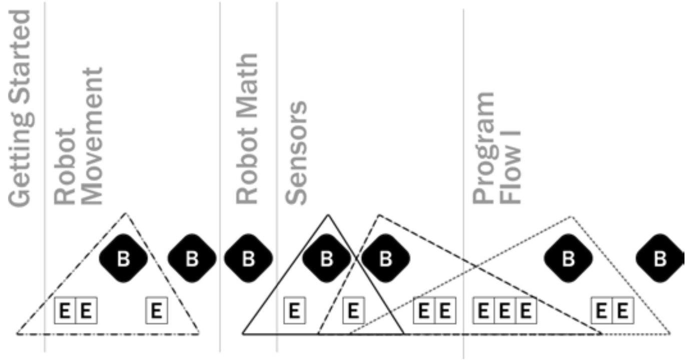


Figure 12. Data collection points across the curriculum.

E = Engagement measured after a mini-challenge. B = Perceived Badge Relevance measured after a chapter challenge. Triangles represent EBE Tris; BEB Tris are not shown, but are simply the reciprocal pattern.

4.2.5.3 RQ1a (Longitudinality)

To address our first research question, whether there is evidence of a longitudinal effect of PBR on Engagement, we conduct regression analyses on the full sample of EBE Tris to test whether a learner's PBR at a given point in time *t* significantly predicts that user's subsequent Engagement at time *t+1*, controlling for the same learner's previous Engagement at time *t-1*. The estimate of the effect of PBR_t on $Engagement_{t+1}$ controlling for $Engagement_{t-1}$ captures learners' relative change in engagement predicted by their relative levels of PBR – that is, whether students with higher PBR tended to shift upward or downward in engagement relative to students with lower PBR.

We model this relationship three different ways in order to establish convergent results. First, we use OLS multiple regression, collapsing multi-item variables to their mean scores. The second method uses a Structural Equation Model in which PBR is treated as a latent factor measured by its indicator items – this technique combines the measurement and structural models into a single step and allows us to test the fit of the model to the observed data. It also retains more of the data using full information maximum likelihood to deal with missing data rather than implementing listwise deletion. The third method uses multi-level Structural Equation Modeling to "control for" overall differences between the different class groups in our study, e.g., if a classroom with a particularly interesting instructor has higher engagement ratings overall. For brevity, we report only the final model here, and provide details of the other analyses in Appendix B. All analyses were performed using Mplus Version 8 and maximum likelihood estimation with robust standard errors (Muthen & Muthen, 2017).

Finally, we test the final model on subsets of data from different portions of the curriculum (looking only at the Tri from Chapter 1, only at the Tri from Chapter 2, etc.). This has two purposes: it allows comparison in a subset of the data in which individual student-level nesting may be safely ignored (rather than left unmodeled out of necessity); and it tests our assumption that Tris are comparable across the curriculum.

4.2.5.4 RQ1b (Demographics and Motivation)

To refine our model and test whether motivational confounds may be responsible for some of our observed relationships, we analyze the subset of data for which motivational pre-survey data is available. We verify that the final model from RQ1a is still a good fit for this subsample, then introduce motivational covariates from the pre-survey as predictors of both initial status and change, to see whether they predict change in engagement in lieu of the PBR predictor (i.e., “explain away” its covariance). If they do not, then it suggests that PBR retains a unique relationship to relative growth in engagement above and beyond those motivational factors.

4.2.5.5 RQ2 (Equity)

Our second research question concerns the for-whom question about badging effects. Again, using the subsample of the original data for which pre-survey data was available, we test interaction effects to see whether the predictiveness of PBR on Engagement varies significantly by age, gender, or minoritized race status. Interactions with latent variables are relatively new in Structural Equation Modeling. Mplus implements a technique called latent moderated structural equations, in which interactions with latent constructs (such as PBR) are modeled as random effects. This technique does not produce conventional fit statistics, and so its fit cannot be evaluated in the usual way. Instead, we use the information criteria (Aikake’s Information Criterion and the Bayesian Information Criterion) and ratio of log-likelihoods (Maslowsky, Jager, & Hemken, 2014) to compare the information efficiency of the model containing the interaction, to the model containing only the main effect of that predictor. The χ^2 difference test of log-likelihoods will be significant if the interaction model is a significantly better fit to the data than the model without the interaction. A failure to reject the null hypothesis implies that the interaction model is a comparable or worse fit. Continuous predictors such as age are grand mean-centered for this analysis.

Additionally, while we would like to test these interaction effects with all motivational and demographic factors simultaneously, the complexity of models increases rapidly in the presence of latent interactions, and thus model nonidentification becomes an issue. For this reason, we test only one demographic factor in each model, using a separate model for each factor. The contrast model is illustrated in Figure 13, using dotted lines to identify the portion that is different between the models.

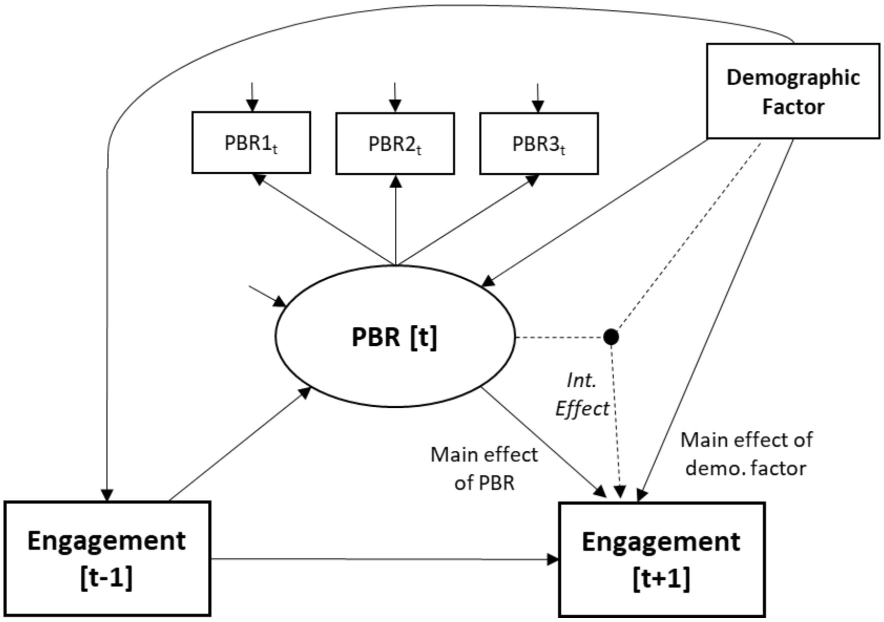


Figure 13. Diagram of the models used to test interaction between each demographic factor and the PBR effect on engagement.

In the main effect model, the interaction effect (dotted line portion) is not present.

4.2.5.6 RQ3 (Reciprocality)

Our final research question is whether we see evidence of reciprocal effects in which higher engagement predicts a relative rise in Perceived Badge Relevance. This analysis uses the full set of CS2N responses as in the RQ1 analyses, but examines the reciprocal set of “BEB” Tris. Using a similar model-building approach as before, we examine whether higher levels of engagement predict relative change in perceptions of badge relevance.

4.2.6 Sample

Our sample includes all CS2N users enrolled in a class Group who completed at least one activity in the selected programming curriculum module between June 2017 and March 2018, and had not opted out of research data collection. The Group membership constraint means that our sample is generally constrained to users operating in classrooms or organized clubs (rather than independent) settings. As our basic unit of analysis is the Tri, our sample is also effectively constrained to those users who completed enough consecutive activities to form Tris, excluding users who completed too few activities, or skipped around too much to generate sufficient adjacent activity data. Ultimately, we collected n=3,696 EBE Tris from 2,410 users in 189 groups, and n=6,863 BEB Tris from 2,980 users in 236 groups.

Our pre-survey data comes from students of 10 teachers recruited from a summer teacher professional development for the curriculum. Not all students who filled out a pre-survey could be matched to their CS2N activity data. This is largely due to students mis-entering their user names on the survey. The subsample for which we were able to match pre-surveys (interest, identity, and demographics) to CS2N activity surveys (engagement and PBR) included 1,832 EBE Tris from 458 users in 38 groups.

4.3 RESULTS

4.3.1 Full Sample Analysis

4.3.1.1 Descriptive statistics

We began by examining the EBE Tri data. A list of means, standard deviations, and bivariate correlations are shown below. Skewness and kurtosis are less than ±1.0 for all variables except PBR3, which some signs of being at-ceiling. This could result in a slightly restricted range, which will bias our results slightly toward non-significance.

Table 3. Means, SDs, and Correlation Matrix for EBE Tris in the Full Data

		Non-nested				Within-Groups (nested)			
	Mean (SD)	Eng_{t-1}	$PBR1_t$	$PBR2_t$	$PBR3_t$	Eng_{t-1}	$PBR1_t$	$PBR2_t$	$PBR3_t$
Eng_{t-1}	2.74 (0.80)								
$PBR1_t$	3.14 (0.97)	.37				.31			
$PBR2_t$	2.98 (1.04)	.45	.66			.39	.64		
$PBR3_t$	3.21 (0.97)	.42	.66	.71		.37	.63	.68	
Eng_{t+1}	2.71 (0.84)	.69	.35	.42	.38	.64	.29	.36	.32

4.3.1.2 Final model

The OLS multiple regression, single-level SEM, and multi-level SEM analyses achieved convergent results (see Table 3) regarding the central research question of whether PBR predicts relative change in Engagement among Tris. We selected the single-level SEM model as our final model. We regard it as superior to the OLS model because SEM accounts for the measurement model (i.e., the loadings and combination of the PBR items into the latent PBR factor, rather than a simple mean score), and because it provides better missing data handling through full information maximum likelihood. We prefer the single-level SEM model because the data contained relatively little variance at the between-groups level (intraclass correlations for all variables were around .10), and the results did not differ substantively from those of the single-level model; we thus opted for the greater ease of interpretation of a single-level model.

Our final model is shown in Figure 14. The model exhibited good fit with n=3,696 EBE Tris: RMSEA = .04 <.08, CFI = .99 > .95, SRMR = .01 < .06 (Hu & Bentler, 1999). All three PBR indicator items have high loadings on the latent factor (λ = .78, .85, and .84). The relationship of PBR_t with Engagement_{t+1}, controlling for prior engagement (Engagement_{t-1}), is β = .16, interpretable as a tendency for slight upward rank-order shifts in engagement among learners with higher PBR over time; a one standard deviation difference in PBR predicts a 0.16 standard deviation difference in subsequent-timepoint engagement, controlling for previous-timepoint engagement. Unstandardized, a one-point increase in PBR predicts a 0.18-point increase in subsequent engagement. This effect is significant at the *p*<.001 level (*t*=9.66). The model explains pseudo-R^2 = 51% of variance in Engagement_{t+1} readings.

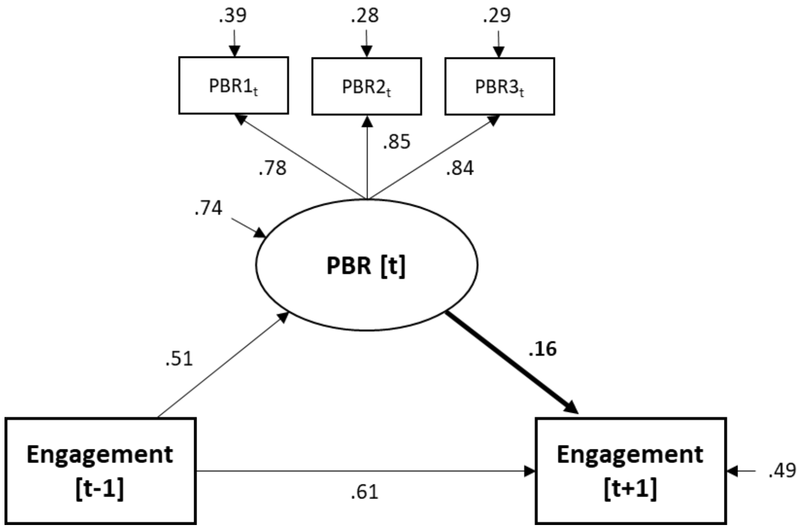


Figure 14. Diagram of final model.

Path weights indicate standardized βs. All values are significant at *p* < .001.

Model integrity checks. To verify that this model is applicable across the various portions of the curriculum, we re-ran the model using only the subset of EBE Tris that corresponded to individual curriculum chunks (see Appendix C). All Tri subsets achieved good fit and a significant positive effect of PBR on Engagement. To rule out significance inflation from the "collapse" of all the Tris into a single pool, we verified that the estimated effect size for the whole set would have remained significant even using the largest standard error in any single-lesson subset. The effect remained significant. We also replicated our results with listwise deletion, instead of full information maximum likelihood estimation.

Table 4. Comparison of PBR_t Effect Estimates from Different Methods

	OLS (scale means)	SEM	MSEM (nesting within classrooms)
PBR_t → Eng_{t+1}	.14	.16	.14
Eng_{t-1} → PBR_t	*Not modeled*	.51	.45
Eng_{t-1} → Eng_{t+1}	.63	.61	.58

4.3.2 Subsample Analysis

4.3.2.1 Descriptive statistics

There were 1,832 EBE Tris in the subsample for which demographic and motivational pre-survey data was available. The means, standard deviations, and Pearson correlations are shown in Table 5.

Table 5. Means, SDs, and Correlation Matrix for EBE Tris in the Pre-Survey Matched Data Subset

	Mean (SD)	Eng_{t-1}	PBR_t	Eng_{t+1}	Identity	Interest	Age	Female
Eng_{t-1}	2.87 (0.86)							
PBR_t	3.25 (0.92)	.44						
Eng_{t+1}	2.80 (0.90)	.80	.44					
Identity	2.06 (0.73)	.26	.45	.23				
Interest	2.40 (0.75)	.40	.28	.33	.66			
Age	12.6 (1.01)	-.22	-.10	-.19	-.10	-.18		
Female	49%	-.12	-.04	-.09	-.28	-.31	.04	
Minoritized	12%	-.02	.05	-.06	.08	.14	.09	-.14

Multi-item scales are collapsed to their means here, but were modeled separately using latent factors. See Appendix D for item-level detail.

Compared to the full sample, the sub-sample had higher mean responses on PBR items, and accordingly, the ceiling effect on PBR item 3 (“I want to earn the badges in this program”) increased. Correlations between the Engagement and PBR factors resemble those in the full sample. Motivational items were highly correlated with each other and moderately correlated with Engagement, but only slightly correlated with PBR. Age has relatively low variance (SD=1.0 years), and thus relatively low observed correlations with the other factors. Sex has a small negative correlation with engagement, but is more strongly correlated with low programming interest and identity. Interest and identity had fairly high correlations (around .6) with each other’s items, suggesting that there would be some danger of multicollinearity when including both constructs in the same model.

4.3.2.2 Motivational predictors

We first verified that the final model from the previous analysis had good fit to the pre-survey matched subset of the data. The fit was acceptable with n=681: RMSEA = .07 <.08, CFI = .99 > .95, SRMR = .021 < .06 . The standardized regression coefficient of PBR_t on Engagement_{t+1}

(controlling for Engagement$_{t-1}$) was β = .22, with pseudo-R^2 = 65% of Engagement$_{t+1}$ variance explained.

The model including interest, identity, age, and minoritized race status (n=1,832) initially had only marginal fit. Upon examination, the motivational factor Identity did not have any significant prediction of any key factors, but did have a high correlation with Interest (*r* = .72), so it was removed on suspicion of causing multicollinearity issues during estimation. The resulting model, shown in Figure 15, had good fit (RMSEA = .04 < .08, CFI = .98 > .95, SRMR = .03 < .06). An analogous model which included Identity but not Interest, produced substantively identical results (RMSEA = .05, CFI = .97, SRMR = .03, Identity factor loadings > .7; Identity predicted the same constructs as Interest, did not change the significance of other predictors, and produced the same estimate for the PBR effect: β = .21 vs. β = .20). We therefore describe only the Interest-based model moving forward.

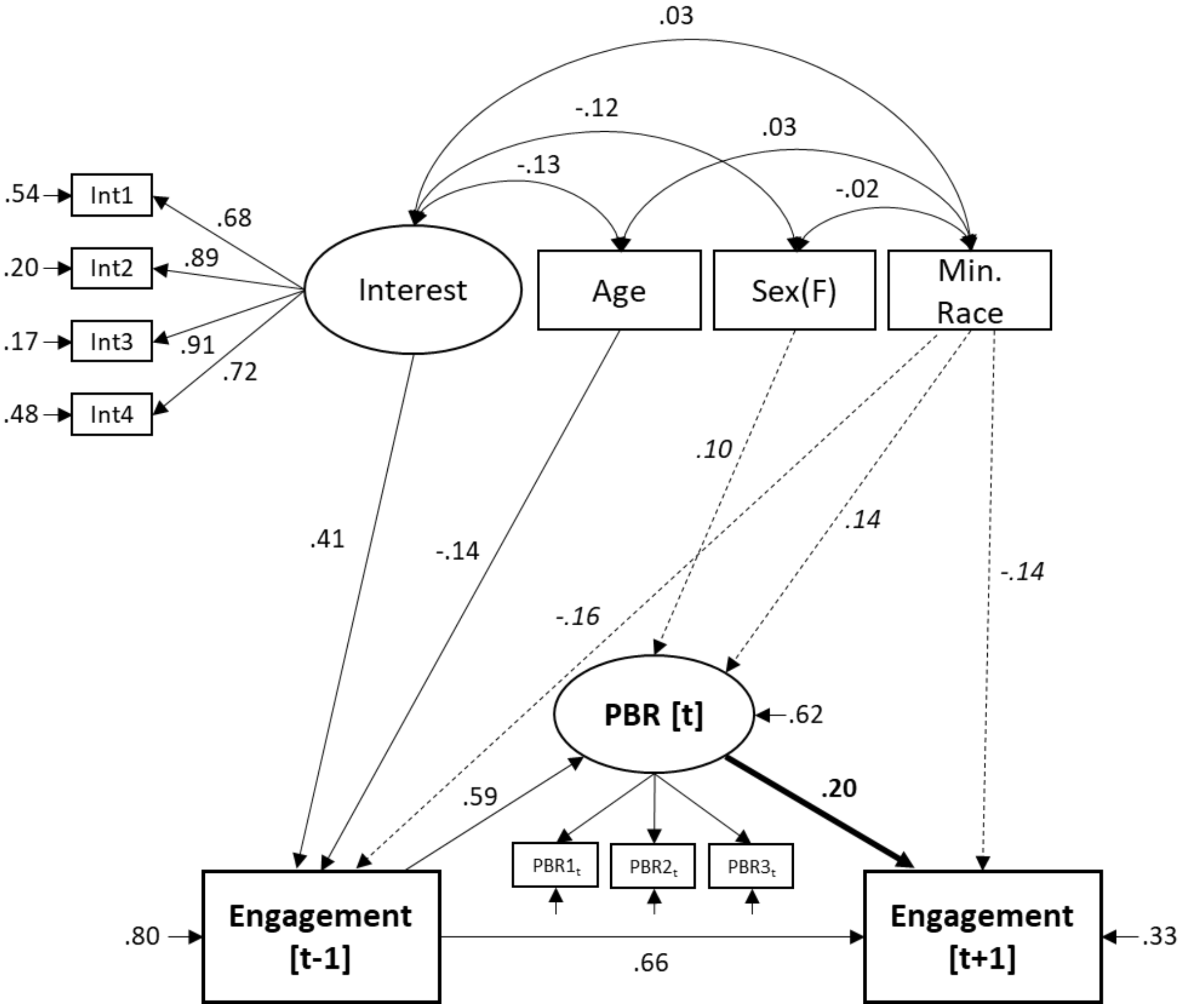


Figure 15. Model diagram for the final demographic model.

All demographic and motivational predictors were estimated as predictors of Engagement$_{t-1}$, PBR$_t$, and Engagement$_{t+1}$, but only significant links are shown. Dotted lines represent marginally significant relationships. Standardized coefficients are shown for latent and continuous factors, unstandardized coefficients are shown for categorical factors.

Motivational and demographic predictors had many expected, and some small unexpected correlations with each other: girls had lower interest (r = -.34, p < .001), but students with minoritized racial background had higher interest (r = .13, p < .001). Girls with minoritized racial background were less common in the sample overall (r = -.14, p < .001), and students with minoritized racial backgrounds tended to be slightly younger (r = .09, p < .001).

Prior-timepoint engagement values, Engagement_{t-1}, were significantly predicted by interest (β = .41, p < .001) and age (β = -.14, p < .001), but not by sex (p = .72), and marginally by racial minoritization status (β = -.16, p = .08). Perceived Badge Relevance, PBR_t, was predicted by prior-timepoint engagement (β = .59, p < .001), but only marginally by sex (β =.10, p = .06) and racial minoritization status (β =.14, p = .11); neither interest (p = .27) nor age (p = .32) were significant direct predictors of PBR. Most importantly, subsequent timepoint engagement – Engagement_{t+1} – was still predicted by PBR_t (β = .20, p < .001) when controlling for prior-timepoint engagement (β = .66, p < .001) and demographics. Among demographic and motivational factors, Engagement_{t+1} was only marginally predicted (negatively) by racial minoritization status (β = -.14, p = .06) and there was no direct prediction of Engagement_{t+1} by interest (p = .41), sex (p = .91), or age (p = .94). Pseudo-R^2 = 67% of variance in Engagement_{t+1} was explained by the model.

Model integrity check. One fundamental assumption of this analysis is that Tris from different portions of the curriculum are comparable enough to group them into a single set. However, when grouped in this way, we ignore the fact that the initial motivational data contained in the pre-survey may be more recent, or less. Moreover, we know from a previous study (Higashi & Schunn, in review) that perceptions of success at lesson activities significantly predict engagement, and share some predictive covariance with PBR. It would not be unreasonable to suspect that motivational changes over the course of the curriculum could be interfering with our analyses, particularly in the later parts of the curriculum. To address the possibility that motivations changed between survey and engagement time points, we verified that our model continued to exhibit good fit and comparable estimates of key relationships in the first Tri alone, i.e., the first lesson used by students. Using the final model with this first-Tri subset of the data (n=458), model fit was still good (RMSEA = .03 < .08, CFI = .99 > .95, SRMR = .03 < .06), the main effect of PBR on Engagement remained significant (β = .36, *p* < .001), and – allowing for the general loss of precision due to a decreased n – only one change in significance occurred: the effect of age on prior-timepoint engagement was negative (β = -.14, *p* < .001) in the full set of Tris, and positive (β = .14, *p* = .013) in the first-Tri subsample. As the number of clusters is the same in the full and first-Tri subsamples (J = 38), this cannot reflect a classroom selection bias – instead, it may mean that older students were more engaged initially, then regressed toward the population mean.

4.3.2.3 Interaction models

To test whether the impact of digital badges appears to be equitable by sex, age, and race, we constructed a series of models that included these factors in interaction with PBR_t. Since models containing random effects do not generate conventional fit statistics, we compare the AIC, sample size adjusted BIC, and log-likelihoods of the interaction model to main effect model to determine whether a random effect is warranted. Comparisons of main effect and interaction models are shown in Table 6.

Table 6. Model Fit and Unstandardized Regression Coefficients for the Effect of Sex, Age, Minoritized Race Status, and the Interactions of Those Factors with PBRt on Engagermentt+1, Controlling for Engagementt-1 and PBRt. ns p>=.10, ~ p<.10, * p<.05, ** p<.01, *** p<.001.

	Sex (female)		Age (years)		Minoritized race status	
	Main only	Interaction	Main only	Interaction	Main only	Interaction
AIC	7180.583	7182.359	7636.169	7637.376	7505.567	7507.495
SSA-BIC	7211.760	7215.176	7668.337	7671.236	7537.376	7540.979
Log-likelihood	-3571.292	-3571.179	-3799.085	-.3798.688	-3733.783	-3733.748
[χ^2 difference test]		p=.63		p=.37		p=.79
PBR main effect	.29***	.28***	.26***	.26***	.27***	.26***
Factor main effect	*ns*	*ns*	*ns*	*ns*	-*.13~*	*ns*
Interaction effect	-	*ns*	-	*ns*	-	*ns*

All three interaction models achieved similar fit to their main effects-only counterparts, with similar AIC and SSA-BIC values and null χ^2 difference test results. In addition, all three models also estimated interaction effect sizes that were statistically indistinguishable from zero. Furthermore, standard errors for the interaction effects were small compared to the effects themselves (e.g., the interaction effect between PBR and sex had a standard error of 0.07 points, but the main effect of PBR itself was estimated at B=0.28 – even at the plus-two-standard-error bound, the difference in effect sizes is only around half the size of the effect itself. Thus, we conclude that the predictive effect of PBR_t on subsequent-timepoint $Engagement_{t+1}$ was not significantly different by sex, age, or minoritized race status within our sample.

4.3.3 Reciprocal Effects Analysis

4.3.3.1 Descriptive statistics

A list of means, standard deviations, and bivariate correlations are shown in Table 5. Skewness and kurtosis are less than ±1.0 for all variables except PBR3, which again shows some signs of being at-ceiling. PBR also had a slight negative skew.

Table 7. Means and Correlations for BEB Tris

	Mean (SD)	$PBR1_{t-1}$	$PBR2_{t-1}$	$PBR3_{t-1}$	Eng_{t}	$PBR1_{t+1}$	$PBR2_{t+1}$
$PBR1_{t-1}$	3.25 (.90)						
$PBR2_{t-1}$	3.10 (.98)	.62					
$PBR3_{t-1}$	3.35 (.89)	.65	.71				
Eng_{t}	2.82 (.90)	.40	.52	.46			
$PBR1_{t+1}$	3.22 (.96)	.54	.49	.52	.45		
$PBR2_{t+1}$	3.06 (1.05)	.49	.67	.58	.57	.68	
$PBR3_{t+1}$	3.24 (1.00)	.54	.55	.59	.53	.68	.75

4.3.3.2 Model

We use a mirrored version of the final model from RQ1 to address the question of reciprocality directly. Our final model is shown in Figure 16. The model exhibited good fit with n=6,863 BEB Tris: RMSEA = .06 < .08, CFI = .98 > .95, SRMR = .02 < .06. All PBR factor loadings were strong (λ between .76 and .88). The effect of $Engagement_{t}$ on PBR_{t+1}, controlling for prior PBR_{t-1}, is $\beta = .24$ ($p < .001$), indicating that learners with higher overall engagement tended to rise in PBR relative to their peers over time. The overall model explains pseudo-R^2 = 54% of variance in PBR_{t+1}.

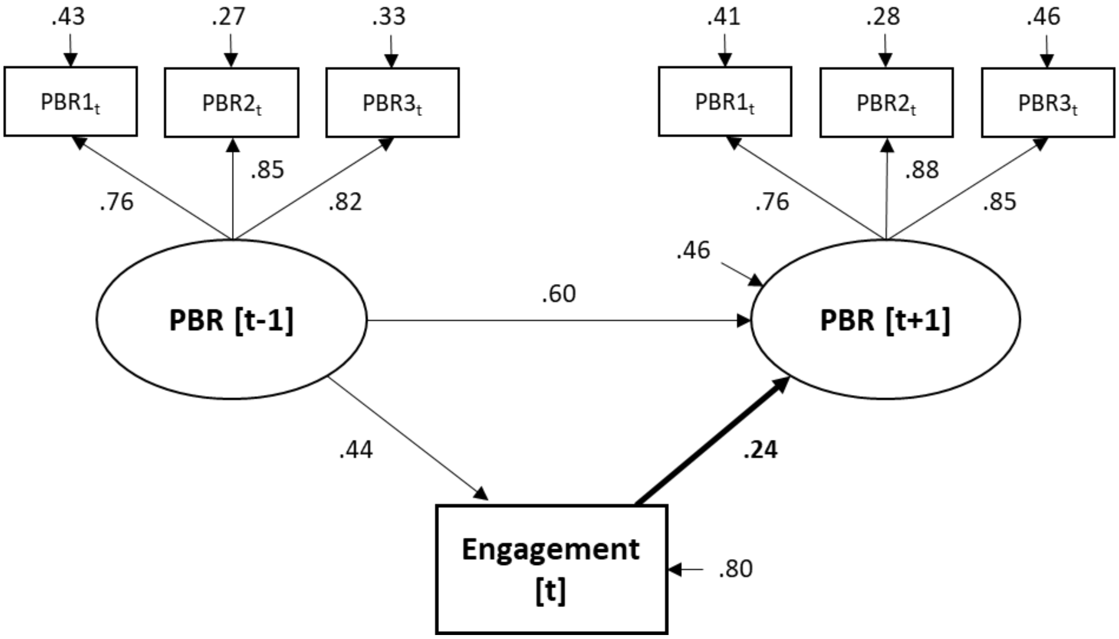


Figure 16. Diagram of final reciprocal (BEB) model.

Path weights indicate standardized βs. All values are significant at $p < .001$.

Model integrity. Model fit remained satisfactory (RMSEA < .08, CFI > .95, SRMR < .06) when running the model on subsets of data belonging to each of the content module Tris separately, and similar effects were found in each case.

4.4 DISCUSSION

In order to inform the theory and design of digital badges for education, we set out to answer three main questions in this study: (RQ1) whether PBR predicts engagement; (RQ2) whether PBR effects appear to be equitable by race, sex, and age; and (RQ3) whether the relationship between PBR and engagement appears to be reciprocal.

4.4.1 Does PBR Predict Engagement?

One major goal of this study was to more rigorously test the relationship between PBR and Engagement: whether the associations observed across programs between PBR and Engagement in a previous study potentially reflect a process involving PBR per se, or whether they were an artifact of other program differences in that cross-sectional dataset. The finding of a positive association was replicated, albeit not of directly comparable magnitude. The previous study estimated a standardized regression coefficient of β=.37 between overall PBR and overall Engagement, between individuals nested in programs, and controlling for overall age, program size, and individual levels of perceived success. The present study estimates a standardized regression coefficient of β=.17 between PBR_t and Eng_{t+1} when controlling for previous-timepoint Engagement. The relationship with PBR is thus small compared to the much larger stability of engagement over time (β = .61).

That the effect size estimate is smaller than in our previous study may seem disappointing at first, but structurally, a previous-timepoint value of the outcome construct is a much stronger control – this is immediately apparent as the models in the current study account for 50-66% of variance in engagement, whereas only 37% of variance in engagement was explained in the prior study (Higashi & Schunn, in review). This stronger control accounts for, e.g., idiosyncratic individual-level measurement factors, as well as unobserved baseline factors that might exert a constant pressure on Engagement. The interpretation of the PBR_t-on-$Engagement_{t+1}$ effect is stronger accordingly: it is the degree to which PBR predicts rank-order shifts in Engagement. That is, the β=.16 effect describes the predictive strength of PBR in picking out episodes in which learners are becoming more engaged over time. This formulation of the PBR-on-Engagement effect is thus smaller in magnitude than previous estimates, but also surer in substance, and reflects an incremental effect at each time point rather than the cumulative results of PBR on engagement across a whole program. It is therefore a substantively better estimate of the extent to which perceived badge relevance is indeed behaving as a receptiveness to badge effects over time, as opposed to a simple cross-sectional correlate.

This case is further strengthened by the fact that controlling for motivational and demographic characteristics did not change the estimate. In the smaller subsample for which we had motivational and demographic covariates available, the PBR_t-to-Engagement_{t+1} effect estimate was β=.22 with no controls, and β=.20 after the addition of interest, age, sex, and minoritized race status. These covariates predicted initial engagement as expected, but did not "explain away" the relationship with PBR as an illusory effect of correlation between PBR and other forms of motivation—the relationship with PBR remained distinct.

4.4.1.1 Contribution to theory

Our finding supports an "ongoing process" interpretation of learners' relationship with digital badges. Looking at particular time-slices in which a learner has a particular value of PBR, we find that in episodes where the learner has higher PBR, that learner tends toward higher subsequent engagement, compared to episodes where observed learners had lower PBR. These effects were observed over a great many episodes drawn from across the multi-week curriculum, which suggests that an active, ongoing process is at work, rather than a one-time evaluation or a static baseline propensity (especially since it also persisted after accounting for motivational and demographic factors). This aligns with the theoretical stance that badges function the way other contingent rewards do: that they are continually interpreted and evaluated by recipients based on their own complex relationship to the topic of interest, who then act accordingly (Deci, Koestner, & Ryan, 2001; Ryan & Deci, 2000). However, neither general domain interest (a relatively content-local covariate) nor age, race, and sex (social covariates linked with a broad range of high-level effects) explained away the observed effects between badge perceptions and engagement – in fact, less than one-tenth of the estimated effect is lost (β=.22 vs. β=.20) with the addition of all four covariates. This suggests that at least some of the observed effect may be unique to badges, or at least processes which are closely tied to individuals' ongoing interpretation of badges. Such an effect is consistent with our theory that learners' perceptions of badges' relevance– often manifested in qualitative studies as reports of badges having value, desirability, contextual meaning, or non-triviality (Suhr, 2014; Davis & Singh, 2015; Wardrip et al., 2016) – play a role in learner decisions to engage with the badged activities, and continue to do so over time.

4.4.2 Are the Effects of PBR Equitable?

Regarding the equity of badge effects, the story is somewhat more complex. The direct relationship of PBR with relative change in Engagement did not differ by age, sex, or racial minoritization status. This means that students who express the same regard for the badges' relevance will tend toward similar shifts in engagement.

However, not all groups of students may have been equally predisposed toward badges in the first place. Such imbalances manifest in two ways: direct relationships between equity factors and PBR, and indirect relationships in which PBR "inherits" an imbalance from engagement. In terms of direct relationships, PBR was marginally predicted by sex (β = .10, *p*=.06) and race (β = .14, *p*=.11). Surprisingly, these leanings favor the two groups that are typically disfavored – PBR is marginally higher for girls than boys, and for minoritized students than White or Asian students. Nevertheless, neither relationship fully passed the threshold of statistical significance, so we more generally conclude that self-reported levels of PBR are similar among learners in each of the equity categories who have the same level of prior engagement. Both PBR and the relationship between PBR and growth in engagement are neutral in terms of direct relationships to equity factors.

The indirect relationships within the model are less exciting, as PBR was strongly predicted (β = .59) by prior-timepoint engagement. This means that PBR will inherit a large proportion of any unequal standing that is already manifested in higher or lower levels of engagement. Unfortunately, prior-timepoint engagement was significantly predicted by interest (β = .41), age (β = -.14), and marginally by minoritized racial status (β = -.06, *p* = .08); girls had lower interest in the course domain (*r* = -.34), and so tended toward lower engagement. Thus, traditionally disfavored groups in STEM – girls and students from minoritized racial backgrounds – come into class with lower engagement, and PBR's neutral direct effects largely carry them forward.

From a theoretical perspective, these results raise other interesting questions. The enticing marginal results around PBR we observed suggest the possibility of badges with a gap-narrowing design via direct impact on perceived relevance. The badges in this study were very closely tied to the curriculum, and did not seem particularly likely to appeal to any group over another on the basis of messaging or aesthetics. Indeed, one of the items in the PBR scale ("I want to earn the badges in this program") was near ceiling, suggesting that these badges were well designed to be broadly appealing to students. Yet if these badges may have possessed weak versions of gap-narrowing elements, what might stronger versions look like – and why?

Ultimately, our mixed equity finding has a substantial implication for both practice and research, in that it suggests that there could be value in continued exploration of the space –

provided we do so with caution. Had we found strong evidence that badges *amplified* the existing biases that lead to inequitable representation in technology occupations, i.e., “made the rich richer”, it might have put a considerable ethical damper on future research. This was not the case, but neither did we find strong evidence of badges counteracting pre-existing inequities.

4.4.3 Does Engagement Also Predict Subsequent PBR?

At first glance, this finding may be somewhat surprising. While it’s intuitive to understand that PBR reflects learner receptiveness to badge effects, it may not be not obvious that students who more actively participate in, think about, and get excited when doing computer programming subsequently start to see programming *badges* as more valuable, desirable, or meaningful. Note that domain interest had no direct predictive effect on PBR in our EBE model, so the mechanism would not seem to be simply becoming more interested in computer programming. Yet, not only is the relationship significant (β = .24, SE=.02), it is arguably larger than the relationship between PBR and engagement (β = .16, SE=.02).

This pattern of results is, however, quite in line with the patterns described by Lave and Wenger’s (1991) theory of Legitimate Peripheral Participation (LPP) in a community of practice, where engagement in disciplinary activity leads a novice to become more attached to the customs, norms, and practices of the discipline. The earning of program-embedded badges squares well with the notions of disciplinary customs and norms.

Of course, a critical outcome of LPP is a tendency toward full participation in the community. Our findings are now consistent with this phenomenon as well – the combination of a PBR-to-Engagement effect and an Engagement-to-PBR effect forms a positive feedback loop, in which the two factors mutually reinforce each other over time. Students who see the badges as relevant to them are more likely to engage with program materials, which in turn predicts that they will see the badges as even more relevant in the future. These feedback cycles can theoretically cause even small effects to compound into large effects over time. This raises the possibility that even if badges have only a relatively weak relationship to engagement, they may yet be impactful. This idea of a compounding effect over time merits additional attention by both researchers and practitioners – some of badges' best effects may only be observed by their consistent application with the same cohort over longer periods of time.

Finally, the feedback loop theory also raises a potentially valuable point regarding badged interventions – if there is indeed a cyclical mechanism in effect, then either badges *or* engagement may serve as effective on-ramps to that cycle. In fact, given the relative effect sizes, it may be more effective to attempt to increase learner engagement first, using badges to strengthen and compound the effect over time. This is a promising and practically testable approach for both research and practice.

4.4.4 Limitations and Future Directions

While this analysis extends our previous findings from cross-sectional to longitudinal, we caution that we have still not yet strongly established a causal relationship between badges, learners' perceptions of badges, and engagement outcomes. A fully experimental or quasi-experimental study design will be necessary to fully address concerns of endogenous bias or unmeasured exogenous influences. Similarly, due to the alternating "Tri" data structure upon which our analysis is based, we cannot entirely rule out the possibility of a concurrent endogenous factor influencing both PBR and Engagement over time. Consequently, one logical direction for future research on perceptions of badge relevance would be random-assignment or other quasi-experimental studies in which fully-known variance in PBR is induced directly (perhaps through the use of different badge systems, or different given explanations) in order to better examine causal hypotheses. Another would be to measure and examine additional concurrent motivational processes, although such a study would need to be careful not to overburden users, whom we have found to be sensitive about classroom time.

There were also environmental factors that we were not able to directly address with this data set. For instance, there could have been substantive differences in the way badges were positioned within different learning environments. Computer programming badges – and robotics programming activities in general – may be regarded differently when they are presented in a mandatory vocational technology sequence, compared to the onboarding process for a robotics competition team. Teachers and parents could also play a role in shaping students' understanding of badges: we have observed some teachers describing badges to students as a simple measure of course progress, while others display them publicly to facilitate competition among students. These qualitative differences in presentation and framing may produce differences in behavior and thinking that overlap in only limited ways with our PBR measure.

Another area we could not examine is the long-term trajectory of badge effects. Our dataset did not include enough data points on individual users, over a long enough period of time to address the question of whether the badge-related relationships with engagement we have observed are novelty effects that wear off, accumulative effects that build up, or both (perhaps resulting in a U-shaped curve). Longer-period datasets may additionally be able to discern distinct trajectories of badge use and integration through the use of mixture modeling or ethnographic methods.

Finally, although our previous findings suggest that the relationship between PBR and engagement is similar across programs, we also acknowledge that this study took place within a single digital badging system and content domain. The badges used in this study are anchored very firmly to curricular progress, in both wording and function. This contrasts sharply with badge systems that aim to complement traditional curriculum by specifically targeting “soft skills” or skills developed in contexts other than guided instruction. It is not a given that learners’ perceptions of badge relevance would be as strongly tied to program engagement, for badges that are not themselves so obviously tied to program activities. Future investigation should therefore explore how and whether PBR relates to engagement in systems with different badging designs and in different content domains.

5.0 CHAPTER 5

I have presented one design and two empirical studies exploring the potential of modeling learners' individual-level appraisals of badges as an explicit process. I proposed this arrangement as a doubly mediated moderation: Perceived Badge Relevance (PBR) is theorized to be a factor predicting learners' subjective evaluations of a program's badges (mediator 1), which in turn predict the degree to which the learners engage with those badges (mediator 2). Engagement with badges in behavioral, cognitive, and affective ways moderates the impact the badges have on engagement with program activities. From this perspective, PBR acts as a distal moderator of badge effects, and can be interpreted as receptivity to the effects of badges. Additionally, changes to learner motivation or domain knowledge may also change the learner's perceptions of the badges in subsequent encounters, thus the effects of PBR may ultimately "loop" back to create an effect on future PBR. This theoretical arrangement is illustrated in Figure 17.

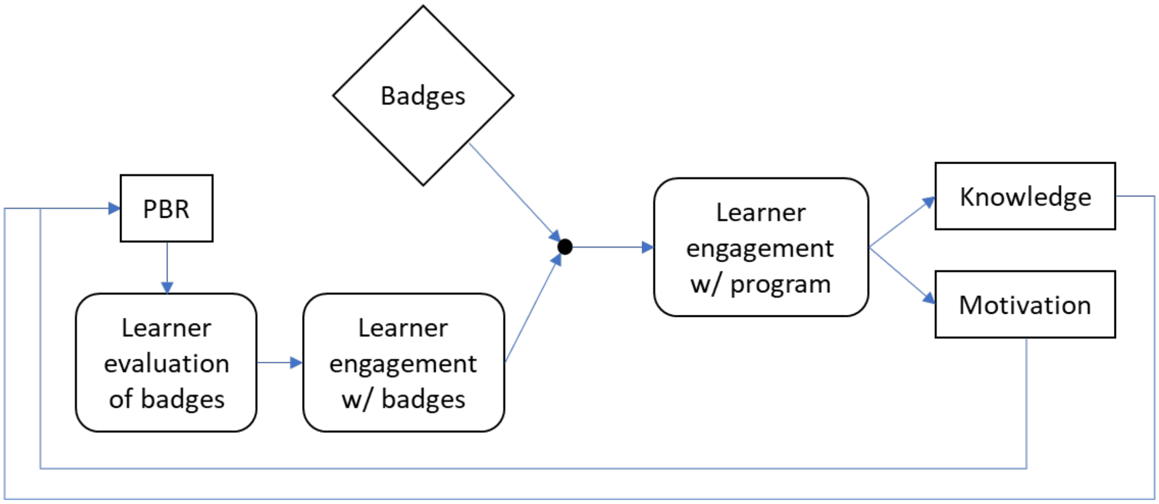

Badges
PBR
Learner evaluation of badges
Learner engagement w/ badges
Learner engagement w/ program
Knowledge
Motivation

Figure 17. Theoretical model including PBR.

The studies presented in Chapters 2 and 4 have focused on establishing empirical support for a simplified version of this model featuring the most theoretically critical relationship – whether there is a detectable effect of PBR on program engagement.

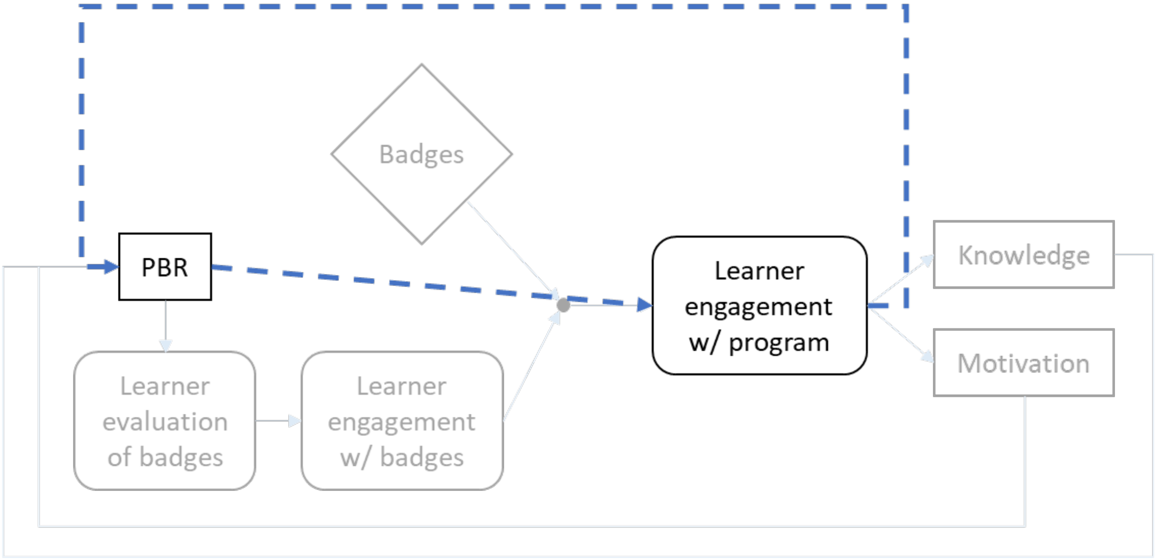

Badges
PBR
Learner evaluation of badges
Learner engagement w/ badges
Learner engagement w/ program
Knowledge
Motivation

Figure 18. Tested portions of the model in these initial studies.

I will arrange the discussion of findings by overarching research question, evaluating the contribution and progress these studies have made toward addressing each question, along with limitations on the presented studies' ability to fully answer them. I will then present recommendations for practice and future research based on the current state of this work, and conclude with a set of next steps specific to my own research agenda.

Note: For convenience throughout this chapter, I will refer to Chapter 2's empirical study of PBR across summer programs as "Study 1", as it is the first empirical study in this dissertation. I will refer to the longitudinal study of PBR in Chapter 4 as "Study 2". Chapter 3's essay on the theoretical design of digital badge evidence systems for assessment will be referred to as the "Evidence essay".

5.1 FINDINGS

5.1.1 RQ I. Support for the Theoretical Model of Action

The most central question in this line of research is whether the process model of badge engagement (see Figure 17) is useful for understanding how digital badges work to impact learners. At minimum, the proposed model must demonstrate two important properties: first, that it has increased predictive power relative to a simpler model that does not include *engagement with badges*; and second, that the predictive power is correctly attributed to the construct of interest (engagement with badges or its proxy, PBR).

5.1.1.1 Is there a predictive effect of PBR on Engagement?

Both empirical studies of PBR identified a positive, statistically significant predictive effect of PBR on engagement in the presence of multiple statistical controls. Study 1 further identified that this effect appeared to express a similar magnitude across programs. Importantly, badges were of varying quality across programs, and indeed learners evaluated them at different levels across programs. It is the relationship between PBR and engagement that was similar across programs. Study 2 used longitudinal controls for prior engagement values to further identify the effect of PBR on changes in engagement over time.

Further, a post-hoc comparison suggests that the addition of PBR explained additional variance. The Study 1 two-level model explained pseudo-R^2=37% of the variance in between-individuals engagement. The same model with PBR removed explained only pseudo-R^2=27%. Under stricter controls and in a longitudinal context, the Study 2 model explained R^2=51% of variance in subsequent-timepoint engagement – with PBR removed, the model explained only R^2=49% of variance. Admittedly, these are not large increases in explained variance, but conceptually PBR is a mediator of the effects of demographic variables and thus the addition of PBR over demographic variables is not likely to greatly increase variance explained. The standardized effect size estimates are also relatively small – β=.36 for the cross-sectional, between-individuals effect in Study 1, and β=.16 for the longitudinal estimate in Study 2.

Returning to the overarching model, however, it is not so surprising that the signal we are able to identify is so weak. According to our model, the effect of PBR must travel through two mediation links and a moderation before having an impact on program engagement. Each of those steps is likely to introduce additional predictive factors, leading to an attenuation of the PBR "signal". For instance, a badge with an attractive graphic may improve learners' subjective evaluation of the badges, without affecting PBR. This second source of variance "dilutes" the overall predictive effect of PBR on subsequent steps in the model. Given the number of steps between PBR and program engagement in the model and the number of potential factors that could influence each, the fact that a detectable effect persists through to the program engagement step suggests that the original effect may actually have been quite strong.

5.1.1.2 Are badges a "process"?

I proposed in Chapter 1 that a shift in thinking might be necessary to understand *how* badges work, predicated on the premise that badges' operation might be better understood as a process involving learner interpretation and evaluation, rather than a black box of inputs and outputs. The overall finding of a significant predictive effect of PBR on engagement in both studies is consistent with this idea – the traditional "black box" explanation, refuted earlier, is that learners' reactions to badges are due to their existing motivational states. Instead, these studies find that learners' subjective opinions about badges have an independent effect on their outcomes. It is difficult to explain the action of an explicitly badge-facing subjective factor without breaking open the black box. It is difficult to explain how – if badge processing is not an ongoing process – the same learners exhibit additional effects of this subjective evaluation factor over time, when controlling for their previous-timepoint values of engagement, as in Study 2. It is especially difficult to explain the "reciprocal" finding in Study 2 that higher levels of engagement increase positive perceptions of badges across time. That badge perceptions are separate from other motivational factors, influential at multiple points in time, and changing over time in an endogenously-influenced way suggest that it is time for the "black box" view to be discarded in favor of models that explicitly account for learners' processing of the badges.

This continual (or repeated) processing model is generally consistent with major theories of motivation, such as Self-Determination Theory's Cognitive Evaluation and Organismic Integration sub-theories (Deci, Koestner, & Ryan, 2001; Ryan & Deci, 2000), in which learners' reactions to external rewards are determined by their individual interpretations of what the reward means in context, based in part on how they themselves feel about their relationship to the domain (e.g., do they identify with it?). Expectancy-Value Theory (Wigfield & Eccles, 2000) similarly suggests that learners perform a sort of unconscious evaluation of whether it is worth it to engage in an activity, as each opportunity presents itself. That badges are also subjectively evaluated by learners as part of a badge use process is therefore not so surprising, but it does need to be modeled.

5.1.1.3 Is the process cyclical?

This question is related to, but separate from, the question of whether badges are a process above. The specific model proposed (see Figure 1) predicts that enhanced motivation and learning caused by increased engagement, in turn induced by badge interactions, will feed back into a learner's interpretive evaluations of the program's badges. In essence, the model predicts a feedback loop. This effect is supported by Study 2, which establishes that both effects – of PBR on subsequent engagement, and engagement on subsequent PBR – will occur over time. These mutually reinforcing effects suggest that learners who are highly engaged, or find the badges highly relevant, will experience accelerating increases over time on both fronts. This effect brings the long-term behavior of the model in line with other established models of motivation such as Expectancy-Value theory (Wigfield & Eccles, 2000), developing individual interest (Hidi & Renninger, 2006; Schiefele, 2009), and legitimate peripheral participation (Lave &

Wenger, 2004); all of which propose recursive models in which motivation and participation mutually reinforce each other over time.

It also opens up interesting possibilities for the use of badges as part of psychosocial interventions like the ones by Cohen et al. (2009), in which a feedback loop is disrupted to prevent an increasingly negative outcome. Badges may be suitable for such uses, but they may also lend themselves to the creation of positive feedback cycles to draw learners further into beneficial learning habits and contexts.

5.1.1.4 Is the effect *really* badges? Is the mechanism really *engagement with badges*?

While the study results have established that the PBR construct has a consistent, longitudinal, and cyclical predictive effect on program engagement, there are separate questions that must be addressed regarding whether (1) the measure accurately and unmistakably represents the construct in the model, and (2) conclusions drawn using our methodology support valid inference about the model as a whole.

The studies presented in this dissertation have provided support for these points, but work in this area is not yet complete. On the one hand, both empirical studies controlled for learner motivational factors – Study 1 controlled for concurrent perceptions of success, which tend to increase learner engagement; Study 2 controlled for initial levels of individual domain interest. Both studies controlled for sex and age. Thus, we can be somewhat certain that the effect we have measured is *not* these factors. Study 2 additionally controlled for previous values of engagement, effectively controlling for a broad array of additional long-term or one-time effects that could have occurred earlier in the learner's experience, as well as overall tendencies for engagement to rise or fall over time. The significant predictive effect of PBR on engagement that persisted suggests that we have identified an effect that, while small in magnitude, is practically separate from domain motivation and a great many other factors.

On the other hand, these are not yet sufficient to establish that the PBR-on-engagement effect is *exactly* what the theoretical model suggests, or that it fully validates the model. The two empirical studies present here have not yet covered enough ground to make that case fully. There are two primary areas that must be covered in future work. The first is straightforward: explicit measures of the remaining constructs (*learner evaluation of badges* and *learner engagement with badges*) must be developed. Without directly observing them, we cannot fully validate the model. The second issue is a matter of an uncontrolled scope confound – while we have controlled for *domain* motivation and some *developmental* and *sociocultural* effects via demographics, we are not strongly controlling for subjective program-level confounds. This means that, even with Study 2's controls for prior-timepoint engagement, the observed variance in responses to the PBR items could be proxying for the learner's general feelings about the *course* at that point in time. Future studies should make sure to control for this potential confound.

Unfortunately, these issues will be difficult to fully address without expanding data collection, i.e., adding more items to our surveys; there is a tricky balance involved in lengthening surveys or increasing the frequency of their collection, as doing so will eventually take a toll on reliability if respondents feel the burden is excessive and begin to supply false answers. I discuss potential ways forward on this issue in the future directions section.

5.1.2 RQ II. Generality of the Model

The theoretical model for the badge-evaluation process is fundamentally contextualized, yet general across contexts. That is, individuals make evaluations of badges that depend on the specific badge system in question, but the process itself should be generalizable to any individual and any badge system.

Study 1 provides some confirmation that this is the case – it uses a two-level model to accommodate between-program differences in the small number of factors that varied significantly between programs, then specifically tested and rejected a random slope hypothesis for the key PBR-to-engagement relationship. This means that the data is better described by a model in which all programs in the sample had the same relationship slope (β=.33) between PBR and Engagement than one in which that slope varied from program to program. That the strength of the PBR-to-Engagement relationship was unchanging from program to program suggests that the relationship may be general across program contexts. Study 2 provided support for generality across time within a program. The longitudinal prediction of subsequent engagement by PBR was present and of similar magnitude for all 4 "Tri" episodes, even though those episodes were drawn from across the multi-week curriculum.

Overall, the two empirical studies have both supported the idea that the model reflects a general process that takes place similarly for all learners.

5.1.3 RQ III. Equity of Badge Processes

As an educational intervention, digital badges must be equitable in their effects. This means that we must observe effects that are either neutral or in opposition to existing biases in the educational system. Particular attention should be paid to age, sex, and race effects. In Study 1, no significant predictive effects of sex or age were found on PBR (socioeconomic status was dropped from the model). In Study 2, no direct prediction of PBR was found by sex, age, or minoritized racial status, and none of these factors had a significant interaction with the PBR-to-engagement relationship when forced into the model for evaluation. In both studies, there was significant prediction of engagement by at least some of these factors.

This means that while PBR itself does not demonstrate any problematic biases, engagement does. Since engagement predicts PBR, this means that PBR may contribute to an *indirect* gap effect, in which the initially lower engagement of, e.g., female students in STEM classes, predicts lower relative PBR, which predicts relatively lower future engagement, and so on. Unfortunately, this is an issue that accompanies every otherwise-neutral positive effect. However, the fact that the problem does not originate in PBR itself, yet PBR has an effect on engagement, means that both engagement and PBR may be good targets for future interventions.

The strength of evidence from Study 2 is slightly imperfect on this topic, as there is a possibility that effects of sex, race, or age might have been missed due to ceiling effects on the PBR measure. Practically speaking, the likelihood of this is very low, as the ceiling effect was not strongly pronounced. Nevertheless, this is a concern that should be more fully addressed moving forward.

5.1.4 RQ IV. Badge System Design Takeaways

The design of digital badges is a broad space with many simultaneous design goals. Chapter 3 of this dissertation unpacked the topic of assessment as it relates to the evidence-bearing function of badges. Attention to use cases, as well as the argumentation structure underlying the evidentiary claim, resulted in the synthesis of four complementary evidence types: eXPeriences, artifacts, endorsements, and exams. These four types of evidence allow consistent evidentiary arguments to be made within several different ontological frameworks, working toward a goal of convergent validation – or, since many skills are "closer" to some ontological traditions than others, to select the best evidence for the job.

Chapter 4 (Study 2) reflected the combination of this theoretical evidentiary architecture with multiple rounds of implementation and testing with real world users – thus, an epistemologically sound environment in which to conduct testing of badges, according to Design-Based Research methodology (Barab & Squire, 2004). Deliberate design choices based on iterative refinement of the badging environment included spreading evidence collection points out among the lessons because users were unwilling to dedicate large chunks of time to specifically "apply" for badges. Naming of badges, lessons, and lesson topics were unified to reduce information load and confusion. This environment was instrumented with a survey engine to collect subjective user experience data that could be easily connected to their curricular progress. We believe that this combination of technical platform elements creates an excellent environment for the collection of ecologically valid data. However, it is not without areas for further improvement, which we will discuss in the future directions section.

Finally, some of our theoretical findings may inform design priorities for digital badging systems. Learners evaluate badges frequently, and their mental evaluations of those badges seem to affect their level of engagement with the lesson activities. Higher levels of engagement also predict increased perception of the system's badges as relevant, so designers of badged content must keep both badge and content quality high. Additionally, the theoretical model (see Figure 1) has unpacked a number of key constructs that badge system designers may wish to instrument for data collection: participants' *perceptions of that system's badges' relevance* and their *engagement with that system's badges*, as well as more traditional motivation and engagement with programs.

5.2 FUTURE DIRECTIONS

I have identified several places where open questions remain in this open line of research. In this section, I will discuss plans to address them moving forward.

5.2.1 Iterating on the PBR Measure

The observed behavior of the PBR measure in two exploratory studies forms the basis for more directed inquiry into factors predicting *engagement with badges*. While the existing three-item PBR measure displays good statistical reliability, and identified a consistent enough "signal" to support our overall theoretical model across multiple programs and longitudinally within a single program, I have also noted several places where the measure has behaved suboptimally.

PBR frequently hovers near-ceiling, which increases the chance of spurious findings (or spurious null findings). As an example, there was a marginally non-significant effect of sex predicting PBR in Study 2, and its direction favored female students for higher PBR. This observation at first seems to be wonderful, if slightly underpowered news. However, this observation is also consistent with the statistical behavior of a system in which *all* group means are increasing, but higher-mean groups have already hit the measure ceiling. Lower-mean groups appear to experience relative gains in this case, because the higher-mean group values are being artificially deflated by the restricted measure range. Thus, the near-ceiling tendency for the PBR measure reduces the certainty with which we can draw inferences about the equitable behavior of digital badges.

Second, the PBR measure could be criticized as being ad-hoc, and to a certain extent, this is a fair criticism. In an attempt to make sure the PBR measure cast a wide enough net that it could reasonably test the subjective badge evaluation hypothesis, the actual items ended up being broad and their precise theoretical alignments unclear.

Therefore, one necessary future direction is to construct a new set of measures for likely predictors of learners' subjective evaluations of badges across badging systems. These items should be deliberately aligned with established theories of motivation, adapted to be *badge*-facing rather than, e.g., domain-facing. This will allow inferences to be made about exactly what *badge motivations* influence the learner evaluation. Further, these measures will need to be calibrated such that they do not risk the same ceiling (or floor) effects that the current measure suffers from.

5.2.2 Next Steps in Model Development

The studies presented in this dissertation tested the simplified two-construct version of the theoretical model (see Figure 2). While this provides encouraging evidence that the full model may also be correct, the evidence is not yet fully compelling. The most important missing pieces, of course, are measures for *learner subjective evaluation of badges* and *learner engagement with badges*. It is imperative that measures for these constructs be measured in order to test the full model. Measures for *engagement with badges* can likely be adapted directly from existing measures of program engagement; measures for *subjective evaluation of badges* may draw from a multitude of motivational theories – expectancy-value theory's (Wigfield & Eccles, 2000) emphasis on subjective evaluation may make it a prime candidate for this purpose. Finally, it is conceivable that *learner evaluation of badges* may become indistinguishable from its motivational predictors, and simply collapse to *learner badge motivation*.

5.2.3 Improving Data Collection Methods

Many of the methodological limitations in Studies 1 and 2 ultimately descend from a tradeoff between the amount of data one can collect from a participant in a sitting, and the willingness of that participant to continue answering questions effortfully and truthfully. In general, user fatigue is a major impediment to data collection around complex constructs. This can result in users responding inattentively to items, or leaving the study altogether. Worse, such attrition tends to be biased – those who leave are likely, for instance, to be more intrinsically motivated by the subject material than those who stay. The remaining participants, and therefore the remaining data, thus overrepresent the portion of the population that is more likely to stay. This is, of course, a major problem for external validity – results from a biased subsample do not generalize to the overall population.

In large-population environments that are relatively common in online learning, however, there may be a solution involving random question sampling from a larger pool of items, from a larger pool of constructs, using a Massively Missing Completely At Random (MMCAR) strategy (Revelle et al., 2016). As missing data has long been a problem in empirical studies, statisticians have developed methods for coping with incomplete data sets. When a data set is missing entries in a completely random way – called Missing Completely At Random (MCAR) – the remaining data can be defensibly used to draw inferences about the overall population, albeit at a penalty to statistical power (as the effective number of subjects is smaller). Data that is missing entries in more systematic ways – perhaps predictable by the remaining data (Missing At Random, MAR), or perhaps not (Missing Not At Random, MNAR) – becomes less and less valid as the missingness becomes more systematic on unobserved factors. Survey fatigue is a case of Missing Not At Random as users systematically drop from the study (or input bad data) in ways that are only partly attributable to their measured aspects of motivation. Conclusions drawn from such a sample are likely to be biased, and the inferences drawn into question.

MMCAR is related to "planned missingness" data collection designs, which attempt to circumvent this threat by deliberately omitting items from each participant's survey in advance. Planned missingness designs do so according to a pre-determined pattern, whereas MMCAR omits (or equivalently, assigns) items at random. This means that the surveys are shorter and less frequent, but as long as the omission pattern is independent of any qualities of the participants, the missing data is not problematically patterned – the data is effectively MCAR. While the attendant reduction in statistical power remains unavoidable, the shorter surveys mitigate the risk of endogenously patterned attrition among participants, and allow the inferences to retain better external validity.

Revelle et al. (2016) have implemented this method in online studies of personality, using random assignment of survey items from a very large pool (several hundred items) with very large numbers of users (hundreds of thousands). The authors' simulations have shown that a manageable reduction to "effective sample size" occurs with MMCAR use, and that a second benefit emerges as well: the use of randomly-assigned items rather than rigidly pre-planned omission matrices (as in, e.g., the conventional "balanced incomplete" design) allow the data collection channel to remain open on an ongoing basis – users can come and go as they please from the website, and a study in need of more statistical power need only leave the door open until a sufficient *n* has been achieved.

Many of the directions for future research I have suggested throughout this chapter inherently involve the collection of more data, from more constructs, and more frequently. Improved data collection methods like MMCAR will be necessary to support that expansion of the theoretical measurement space, without sacrificing the trustworthiness of user responses.

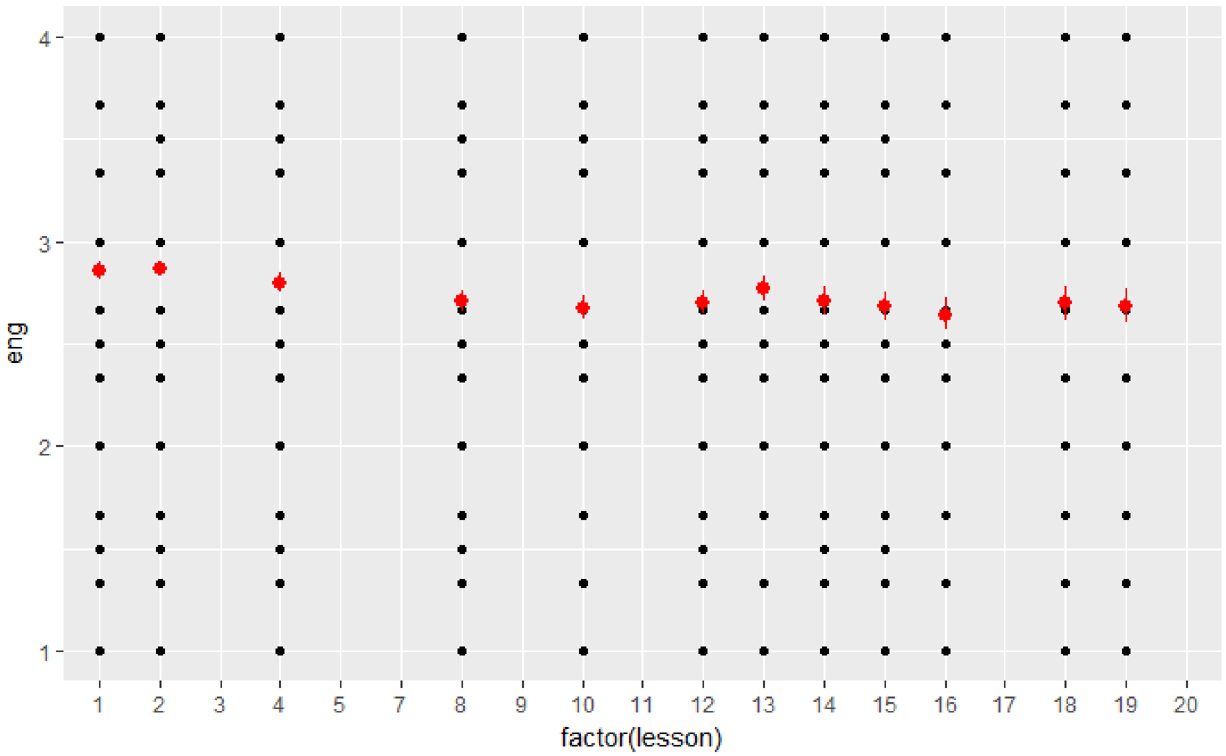
5.3 CONCLUSION

The research presented here represents the early stages of inquiry into a complex topic. Fully understanding the role and constitution of learners' evaluation of and engagement with digital badges is ultimately the same as understanding any other type of motivation – at the end of the day, motivation toward badges is just a special case of motivation more broadly speaking – and just as complex. This does not undercut the substance of the findings presented here, of course. Their major theoretical contribution is to have begun unpacking *how* badges' effects may vary by individual. This research aimed to open up the black box, laying out its contents as an evaluative process. I proposed an initial predictor (PBR) of subjective badge evaluation, and provided exploratory evidence that it predicts a unique portion of program-level engagement, suggesting that the separation of badge perceptions from other motivation is both justified and productive.

APPENDIX A

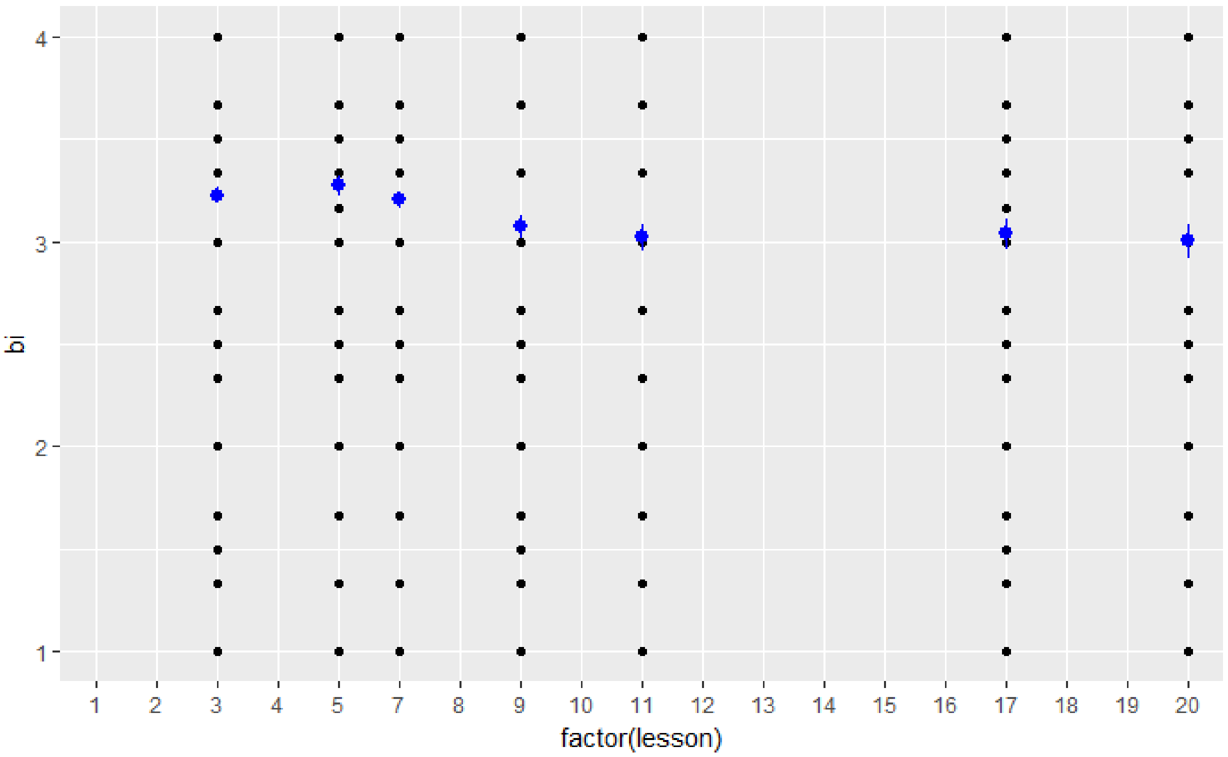
SIMILARITY OF ENGAGEMENT AND PBR MEANS ACROSS LESSONS

A.1 PLOT OT ENGAGEMENT SCORES BY LESSON



Mean scores across Tris for each lesson are in red. Lines indicate bootstrapped 95% confidence intervals for means.

A.2 PLOT OF PBR (CALLED "BI" IN THE DATASET) SCORES BY LESSONS



Mean scores across Tris for each lesson are in blue. Vertical lines indicate bootstrapped 95% confidence intervals for means.

APPENDIX B

ALTERNATIVE ANALYSES FOR EBE TRI FULL SAMPLE

We tested three different methods for modeling the effects of PBR_t on $Engagement_{t+1}$. The final method we selected is described in the Methods section of the main text. Here, we briefly describe the two methods we described as “convergent” in the Results.

B.1 OLS REGRESSION MODEL

We ran a simple OLS multiple regression model of $Engagement_{t+1}$ regressed on PBR_t and $Engagement_{t-1}$. PBR was collapsed to a mean scale score for this version of the analysis (Cronbach’s $\alpha = .85$). The model equation was:

$$Engagement_{t+1} = \beta_1(PBR_t) + \beta_2(Engagement_{t-1}) + \beta_3$$

OLS estimation produces (unstandardized) $\beta_1 = .14$, $\beta_2 = .67$, $\beta_3 = 0.46$ (intercept). Standardized coefficients are $\beta_1 = .14$, $\beta_2 = .63$. Note that this type of regression is unable to simultaneously model the regression of PBR_t on $Engagement_{t-1}$. In a single regression, the standardized coefficient of PBR_t regressed on $Engagement_{t-1}$ is .47 (unstandardized = .52).

B.1.1 Two-Level Structural Equation Model

We also estimated a two-level structural equation model, nesting Tris within classrooms, to account for instructor and other contextual effects. Initial examination of intraclass correlations (mostly around .100) and variance/correlation (small variance with high correlation) in a latent multi-level split suggested that Level 2 (between-classrooms) modeling was unlikely to substantially alter the structure of the Level 1 (between-Tris within classrooms) model. See Table 1.

Table 1. Intraclass correlations and covariance and correlation matrices for latent classroom-level intercepts in the two-level model.

		Covariance					Correlation				
	CC	ng_{t-1}	$BR1_t$	$BR2_t$	$BR3_t$	ng_{t+1}	ng_{t-1}	$BR1_t$	$BR2_t$	$BR3_t$	ng_{t+1}
ng_{t-1}	146	.09									
$BR1_t$	085	.07	.08				.82				
$BR2_t$	116	.09	.10	.13			.85	.943			
$BR3_t$	112	.08	.09	.11	.11		.79	.948	.965		
ng_{t+1}	151	.100	.07	.10	.08	.10	.00	.805	.838	.779	

Additionally, the small ICC and variance caused problems with convergence during estimation. There was insufficient unique covariance between the three indicators items of PBR to model a latent between-groups PBR factor. Consequently, in our two-level model, latent intercepts for each indicator are simply correlated. Additionally, a very high correlation (r=.998) between the latent class-level intercepts of prior-timepoint engagement and subsequent-timepoint engagement dominated estimation at that level – attempting to estimate predictive links between the PBR factors (either together or using a single item as a proxy) yielded only non-significant effects or invalid estimates (e.g., standardized effects > 1.0). The final two-level model is shown in Figure 1.

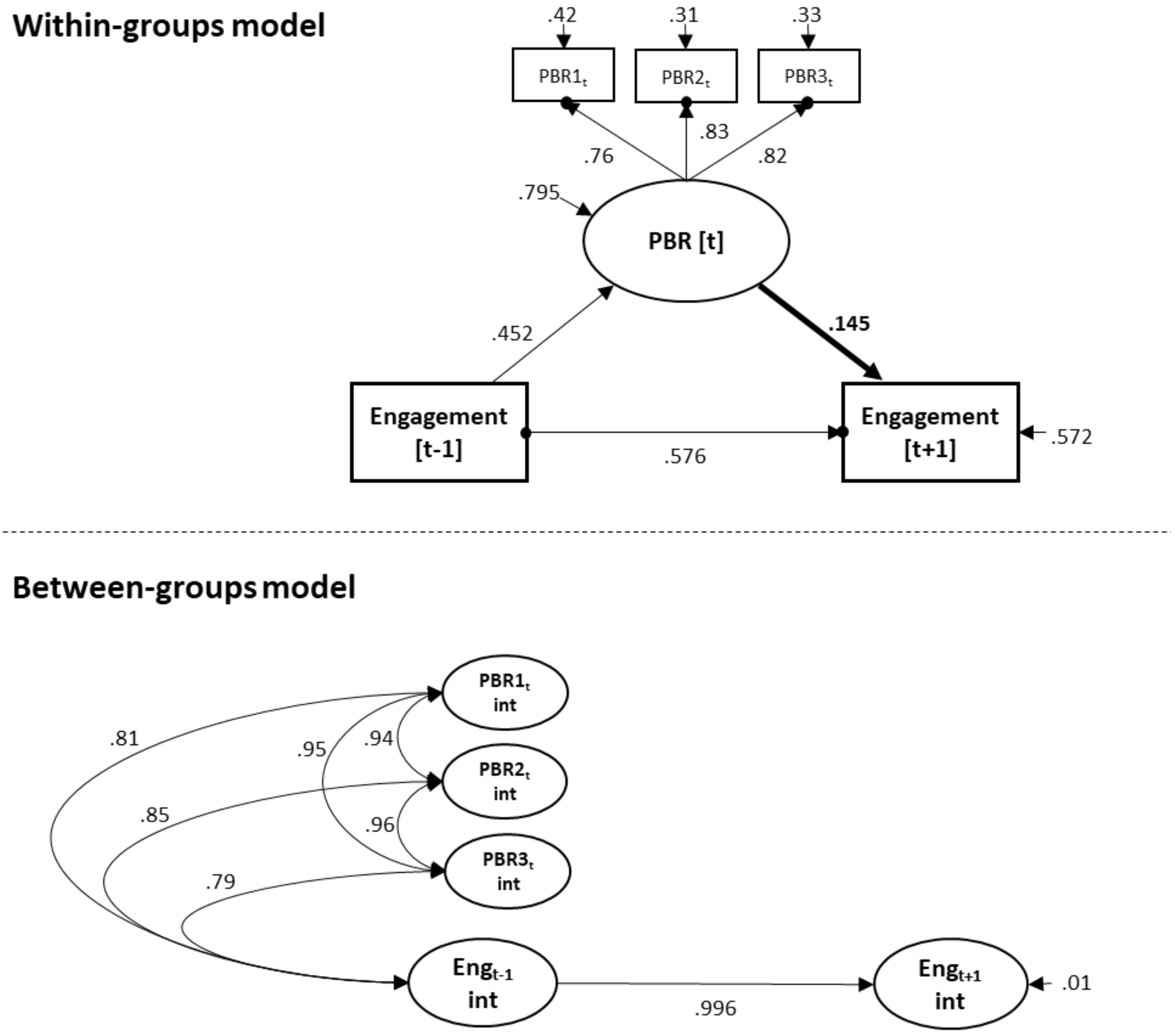


Figure 19. Two-level structural equation model.

APPENDIX C

CURRICULUM MAPPING OF TRIS

C.1 CURRICULUM MAPPINGS OF TRIS

Final TRI name	Segments	Curriculum sections covered
Engagement-Badges-Engagement		
EBE1	1 2 3	E: Robot Movement – Arm + Moving Forward MCs B: Robot Movement – Sensabot E: Robot Movement – Turning in Place MC
EBE2	7 8 9	E: Sensors – Forward Until Near MC B: Sensors – Dynamic Maze E: Sensors – Turn for Angle 2 MC
EBE3	9 10 11	E: Sensors – Turn for Angle 2 MC B: Sensors – Golf Course Mower E: Sensors – Forward Until Red + Traffic Signal MC + Program Flow I – Looped Movements + Loop with Count Control + Loop with Sensor Control
EBE4	11 12 13	E: Sensors – Forward Until Red + Traffic Signal MC + Program Flow I – Looped Movements + Loop with Count Control + Loop with Sensor Control B: Program Flow I – Container Transport E: Program Flow I – Turn if blocked + Looped decision
Badges-Engagement-Badges		
BEB1	2 3 4	B: Robot Movement – Sensabot E: Robot Movement – Turning in Place MC B: Robot Movement – Turning in Place
BEB2	6 7 8	B: Robot Math – Expedition Atlantis Level 1 E: Sensors – Forward Until Near MC B: Sensors – Dynamic Maze
BEB3	8 9 10	B: Sensors – Dynamic Maze E: Sensors – Turn for Angle 2 MC B: Sensors – Golf Course Mower
BEB4	10 11 12	B: Sensors – Golf Course Mower E: Sensors – Forward Until Red + Traffic Signal MC + Program Flow I – Looped Movements + Loop with Count Control + Loop with Sensor Control B: Program Flow I – Container Transport
BEB5	12 13 14	B: Program Flow I – Container Transport E: Program Flow I – Turn if blocked + Looped decision B: Program Flow I – Strawberry Plant Sorter

APPENDIX D

ITEM-LEVEL CORRELATIONS FOR MOTIVATIONAL AND DEMOGRAPHIC ANALYSIS

D.1 MEANS, SDS, AND CORRELATION MATRIX FOR EBE TRIS IN THE PRE-SURVEY MATCHED DATA

	Mean (SD)	Eng$_{t-1}$	PBR1$_{t}$	PBR2$_{t}$	PBR3$_{t}$	Eng$_{t+1}$	ID1	ID2	ID3	ID4	Int1	Int2	Int3	Int4	Age	Fem.
Eng$_{t-1}$	2.87 (0.86)															
PBR1$_{t}$	3.25 (0.92)	.44														
PBR2$_{t}$	3.09 (1.03)	.56	.66													
PBR3$_{t}$	3.33 (0.95)	.53	.70	.72												
Eng$_{t+1}$	2.80 (0.90)	.80	.44	.56	.52											
ID1	2.16 (0.86)	.26	.18	.14	.18	.23										
ID2	2.07 (0.86)	.22	.08	.06	.09	.18	.74									
ID3	1.97 (0.81)	.18	.14	.07	.11	.16	.67	.76								
ID4	2.05 (0.78)	.24	.08	.14	.12	.22	.62	.70	.70							
Int1	2.21 (1.01)	.28	.23	.20	.18	.21	.51	.42	.48	.35						
Int2	2.66 (0.80)	.38	.23	.25	.27	.33	.63	.52	.51	.47	.60					
Int3	2.63 (0.88)	.35	.17	.18	.20	.31	.64	.51	.50	.44	.61	.82				
Int4	2.12 (0.82)	.33	.18	.24	.20	.27	.52	.38	.51	.49	.56	.62	.65			
Age	12.6 (1.01)	-.22	-.10	-.14	-.16	-.19	-.14	-.10	-.10	.01	-.10	-.17	-.15	-.21		
Female	48.7%	-.12	-.04	-.02	-.01	-.09	-.29	-.22	-.26	-.19	-.20	-.31	-.32	-.24	.04	
Minoritized	12.2%	-.02	.05	.03	.01	-.06	.07	.07	.05	.08	.13	.11	10	.12	.09	-.14

BIBLIOGRAPHY

Abramovich, S. (2016). Understanding digital badges in higher education through assessment. *On the Horizon*, *24*(1), 126-131.

Abramovich, S., Higashi, R., Hunkele, T., Schunn, C., & Shoop, R. (2011). An achievement system to increase achievement motivation. Presented at *Games+Learning+Society conference 7*.

Abramovich, S., Schunn, C., & Higashi, R. M. (2013). Are badges useful in education?: It depends upon the type of badge and expertise of learner. *Educational Technology Research and Development*, *61*(2), 217-232.

Activation Lab (2015b). *Measures Technical Brief: Engagement in Science Learning Activities, version 3.1*. In review.

Activation Lab. (2015a). Tools: Measures and Data Collection Instruments. Retrieved September 7, 2015, from http://www.activationlab.org/tools/

Allen-Ramdial, S. A. A., & Campbell, A. G. (2014). Reimagining the pipeline: Advancing STEM diversity, persistence, and success. *BioScience*, *64*(7), 612-618.

American Institute of Certified Public Accountants. (2017) Digital Badges. Retrieved November 08, 2017, from https://www.aicpa.org/cpeandconferences/cpeselfstudy/digital-badges.html

Antin, J., & Churchill, E. F. (2011, May). Badges in social media: A social psychological perspective. In *CHI 2011 Gamification Workshop Proceedings* (pp. 1-4). New York, NY: ACM.

Aschbacher, P. R., Li, E., & Roth, E. J. (2010). Is science me? High school students' identities, participation and aspirations in science, engineering, and medicine. *Journal of Research in Science Teaching*, *47*(5), 564-582. http://doi.org/10.1002/tea.20353.

Asparouhov, T. & Muthén, B. (2006). Constructing covariates in multilevel regression. Mplus Web Notes: No. 11. www.statmodel.com.

Baker, R., Walonoski, J., Heffernan, N., Roll, I., Corbett, A., & Koedinger, K. (2008). Why students engage in" gaming the system" behavior in interactive learning environments. *Journal of Interactive Learning Research*, *19*(2), 185-224.

Baron, R. M., & Kenny, D. A. (1986). The moderator–mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of personality and social psychology, 51*(6), 1173.

Barton, P. E. (2006). *High school reform and work: Facing labor market realities*. Policy Evaluation and Research Center, Policy Information Center, Educational Testing Service.

Bathgate, M. E., & Schunn, C. D. (2017). The psychological characteristics of experiences that influence science motivation and content knowledge. *International Journal of Science Education, 17,* 2402-2432. 10.1080/09500693.2017.1386807.

Beals, A. C., Kazberouk, A., Rosenberg, J., Wachter, K., Choi, S., Yan, Z., & Weintraub, R. (2015). Expanding competency-based credentialing in healthcare: A case for digital badges for global health delivery. *Annals of Global Health*, *81*(1), 71.

Ben-Eliyahu, A., Moore, D., Dorph, R., & Schunn, C. D. (2018). Investigating the multidimensionality of engagement: Affective, behavioral, and cognitive engagement in science across multiple days, activities, and contexts. *Contemporary Educational Psychology, 53,* 87-105.

Bishop, A. R., Berryman, M. A., Wearmouth, J. B., & Peter, M. (2012). Developing an effective education reform model for indigenous and other minoritized students. *School Effectiveness and School Improvement*, *23*(1), 49-70.

Blackburn, R. D., Porto, S. C., & Thompson, J. J. (2016). Competency-based education and the relationship to digital badges. In L. Y. Muilenburg, & Z. Berge (Eds.), *Digital badges in education: Trends, issues, and cases* (pp. 30-38). Routledge, Taylor & Francis.

Book, P. A. (2014). All Hands on Deck: Ten Lessons from Early Adopters of Competency-Based Education. *Western Interstate Commission for Higher Education*.

Boticki, I., Baksa, J., Seow, P., & Looi, C. K. (2015). Usage of a mobile social learning platform with virtual badges in a primary school. *Computers & Education,* 86, 120-136.

Bowen, K., & Thomas, A. (2014). Badges: A common currency for learning. *Change: The Magazine of Higher Learning*, *46*(1), 21-25.

Buchem I. (2016). Digital badges as (parts of) digital portfolios: Design patterns for educational and personal learning practice. In D. Ifenthaler, N. Bellin-Mularski, D. K. Mah (Eds.), *Foundation of digital badges and micro-credentials* (pp. 343-367). Springer, Cham.

Bull, Bernard. (2014) You Can Now Earn a Master’s Degree in #EdTech Through Competency-Based Digital Badges. Retrieved November 07, 2017, from http://etale.org/main/2014/09/07/you-can-now-earn-a-masters-degree-in-edtech-through-competency-based-digital-badges/

Burke, R. J., & Mattis, M. C. (Eds.). (2007). *Women and minorities in science, technology, engineering, and mathematics: Upping the numbers*. Edward Elgar Publishing.

Casilli, C., & Hickey, D. (2016). Transcending conventional credentialing and assessment paradigms with information-rich digital badges. *The Information Society*, *32*(2), 117-129.

Chaiklin, S. (2003). The zone of proximal development in Vygotsky's analysis of learning and instruction. *Vygotsky's educational theory in cultural context*, *1*, 39-64.

Charleer, S., Klerkx, J., Odriozola, S., Luis, J., & Duval, E. (2013, December). Improving awareness and reflection through collaborative, interactive visualizations of badges. In *ARTEL13: Proceedings of the 3rd Workshop on Awareness and Reflection in Technology-Enhanced Learning* (Vol. 1103, pp. 69-81). CEUR-WS.

Chou, C. C., & He, S. J. (2017). The Effectiveness of Digital Badges on Student Online Contributions. *Journal of Educational Computing Research*, *54*(8), 1092-1116.

Cohen, G. L., Garcia, J., Purdie-Vaughns, V., Apfel, N., & Brzustoski, P. (2009). Recursive processes in self-affirmation: Intervening to close the minority achievement gap. *Science*, *324*(5925), 400-403.

Collins, A. (2006). Cognitive apprenticeship. In R. K. Sawyer (Ed.), *The Cambridge handbook of the learning sciences* (pp. 47-60). Cambridge, UK: Cambridge University Press.

Collins, H. M. (2001). What is tacit knowledge. In T. R. Schatzki, K. Knorr Cetina, & E. Von Savigny (Eds.), *The practice turn in contemporary theory* (pp. 107-119). New York, Routledge.

Committee on Underrepresented Groups and the Expansion of the Science and Engineering Workforce Pipeline. (2011). *Expanding Underrepresented Minority Participation: America's Science and Technology Talent at the Crossroads*. Washington, DC: National Academies Press.

CREATE (2015). HASTAC report: Badging & learning. New York, NY: Consortium for Research and Evaluation of Advanced Technology in Education, New York University. Retrieved from http://create.nyu.edu/wordpress/wp-content/uploads/2015/02/HASTAC-Report-Badges-and-Learning-CREATE.pdf

Davis, K., & Singh, S. (2015). Digital badges in afterschool learning: Documenting the perspectives and experiences of students and educators. *Computers & Education*, *88*, 72-83.

Deci, E. L., Koestner, R., & Ryan, R. M. (2001). Extrinsic rewards and intrinsic motivation in education: Reconsidered once again. *Review of Educational Research*, *71*(1), 1-27.

Denny, P. (2013, April). The effect of virtual achievements on student engagement. In *Proceedings of the SIGCHI conference on human factors in computing systems* (pp. 763-772). ACM.

Derryberry, A., Everhart, D., & Knight, E. (2016). In Muilenburg, L. Y., & Berge, Z. L. (Eds.) *Digital badges in education: Trends, issues, and cases*. New York, NY: Routledge.

Deterding, S. (2011, May). Situated motivational affordances of game elements: A conceptual model. In *Gamification: Using game design elements in non-gaming contexts, a workshop at CHI*.

Deterding, S., Dixon, D., Khaled, R., & Nacke, L. (2011, September). From game design elements to gamefulness: defining gamification. In *Proceedings of the 15th international academic MindTrek conference: Envisioning future media environments* (pp. 9-15). ACM.

Deterding, S., Sicart, M., Nacke, L., O'Hara, K., & Dixon, D. (2011, May). Gamification. using game-design elements in non-gaming contexts. In *CHI'11 extended abstracts on human factors in computing systems* (pp. 2425-2428). ACM.

Devedžić, V., & Jovanović, J. (2015). Developing open badges: a comprehensive approach. *Educational Technology Research and Development*, *63*(4), 603-620.

Disessa, A. A., & Sherin, B. L. (1998). What changes in conceptual change?. *International journal of science education*, *20*(10), 1155-1191.

Dorph, R., Cannady, M. A., & Schunn, C. D. (2016). How science learning activation enables success for youth in science learning experiences. *Electronic Journal of Science Education*, *20*(8).

Duncan, A. (2011). Digital badges for learning. Speech given at *4th Annual Launch of the MacArthur Foundation Digital Media and Lifelong Learning Competition*. September 15, 2011.

Duncan, R. G., & Hmelo-Silver, C. E. (2009). Learning progressions: Aligning curriculum, instruction, and assessment. *Journal of Research in Science Teaching*, *46*(6), 606-609.

Durlak, J. A., Weissberg, R. P., & Pachan, M. (2010). A meta-analysis of after-school programs that seek to promote personal and social skills in children and adolescents. *American Journal of Community Psychology*, *45*(3-4), 294-309.

Dziuban C. D., Moskal P. D., Dziuban E. K. (2000). Reactive behavior patterns go online. *Journal of Staff, Program & Organization Development, 17*(3): 171-182.

Eccles, J. (2009). Who am I and what am I going to do with my life? Personal and collective identities as motivators of action. *Educational Psychologist*, *44*(2), 78-89.

Educause. (2017) Badges. Retrieved November 08, 2017, from https://library.educause.edu/topics/teaching-and-learning/badges

Elkordy A. (2016). Development and implementation of digital badges for learning science, technology, engineering and math (STEM) practices in secondary contexts: A pedagogical approach with empirical evidence. In D. Ifenthaler, N. Bellin-Mularski, D. K. Mah (Eds.), *Foundation of digital badges and micro-credentials* (pp. 483-505). Springer, Cham.

Ellis L. E., Nunn S. G., Avella J. T. (2016). Digital badges and micro-credentials: Historical overview, motivational aspects, issues, and challenges. In D. Ifenthaler, N. Bellin-Mularski, D. K. Mah (Eds.), *Foundation of digital badges and micro-credentials* (pp. 3-21). Springer, Cham.

Engle, R. A. (2006). Framing interactions to foster generative learning: A situative explanation of transfer in a community of learners classroom. *The Journal of the Learning Sciences*, 15(4), 451-498.

Falkner, N. J., & Falkner, K. E. (2014, November). Whither, badges? or wither, badges!: a metastudy of badges in computer science education to clarify effects, significance and influence. In *Proceedings of the 14th Koli Calling International Conference on Computing Education Research* (pp. 127-135). ACM.

Fanfarelli, J. R., & McDaniel, R. (2017). Exploring digital badges in university courses: Relationships between quantity, engagement, and performance. *Online Learning*, *21*(2), n2.

Filsecker, M., & Hickey, D. T. (2014). A multilevel analysis of the effects of external rewards on elementary students' motivation, engagement and learning in an educational game. *Computers & Education*, *75*, 136-148.

Finkelstein, J., Knight, E., & Manning, S. (2013). The potential and value of using digital badges for adult learners: Draft for public comment. *American Institute for Research. Retrieved from http://lincs. ed. gov/publications/pdf/AIR_Digital_ Badge_Report_508. pdf* Hamari J., Eranti V. (2011, September). Framework for designing and evaluating game achievements. Proceedings of DiGRA 2011, Hilversum, The Netherlands.

Foli, K. J., Karagory, P., & Kirby, K. (2016). An exploratory study of undergraduate nursing students' perceptions of digital badges. *Journal of Nursing Education*, *55*(11), 640-644.

Foundation for California Community Colleges. (2017) 21st Century Skills Badging. Retrieved November 08, 2017, from https://foundationccc.org/What-We-Do/Workforce-Development/Workforce-Services/21st-Century-Skills-Badging

Fredricks, J. A., Blumenfeld, P. C., & Paris, A. H. (2004). School engagement: Potential of the concept, state of the evidence. *Review of Educational Research*, *74*(1), 59-109.

Garnett, T., & Button, D. (2018). The use of digital badges by undergraduate nursing students: A three-year study. *Nurse Education in Practice*.

Gibson, D., Ostashewski, N., Flintoff, K., Grant, S., & Knight, E. (2015). Digital badges in education. *Education and Information Technologies*, *20*(2), 403-410.

Grant, S. (2016). *Promising Practices of Open Credentials: Five Years of Progress*. Retrieved from https://drive.google.com/file/d/0B7kHRuri9QdPQmRfdXZrblpSX0U/view

Grant, S., & Betts, B. (2013, May). Encouraging user behaviour with achievements: an empirical study. In *Mining Software Repositories (MSR), 2013 10th IEEE Working Conference on* (pp. 65-68). IEEE.

Grant, Sheryl L. 2014. *What counts as learning: Open digital badges for new opportunities.* Irvine, CA: Digital Media and Learning Research Hub.

Gresalfi, M. S. (2009). Taking up opportunities to learn: Constructing dispositions in mathematics classrooms. *The Journal of the Learning Sciences*, *18*(3), 327-369.

Halavais, A. M. (2012). A genealogy of badges: Inherited meaning and monstrous moral hybrids. *Information, Communication & Society*, *15*(3), 354-373.

Hamari, J., & Eranti, V. (2011, September). Framework for Designing and Evaluating Game Achievements. In *Digra conference*.

Hammer, D., Elby, A., Scherr, R. E., & Redish, E. F. (2005). Resources, framing, and transfer. In J.P. Mestre (Ed.), *Transfer of learning from a modern multidisciplinary perspective* (pp. 89-120). Greenwich, CT: IAP.

Hanus, M. D., & Fox, J. (2015). Assessing the effects of gamification in the classroom: A longitudinal study on intrinsic motivation, social comparison, satisfaction, effort, and academic performance. *Computers & Education*, *80*, 152-161.

Hart Research Associates. (2015, January). Falling short? College learning and career success: Selected findings from online surveys of employers and college students conducted on behalf of the Association of American Colleges & Universities. Retrieved from http://www.aacu.org/sites/default/files/files/LEAP/2015employerstudentsurvey.pdf

Heckman, J. J., & Kautz, T. (2012). Hard evidence on soft skills. *Labour Economics*, *19*(4), 451-464.

Hew, K. F., Huang, B., Chu, K. W. S., & Chiu, D. K. (2016). Engaging Asian students through game mechanics: Findings from two experiment studies. *Computers & Education*, *92*, 221-236.

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

Higashi, R., Abramovich, S., Shoop, R., & Schunn, C. (2012). The roles of badges in the computer science student network. *Paper published at the Games+ Learning+ Society conference* (pp. 423-429).

Higashi, R. M., Schunn C., Nguyen, V., & Ososky, S. (2017). Coordinating Evidences Across Learning Modules Using Digital Badges. In R. Sottilare, A. Graesser, X. Hu, and G. Goodwin (Eds.). (2017). Design Recommendations for Intelligent Tutoring Systems: Volume 5 - Assessment Methods. Orlando, FL: U.S. Army Research Laboratory. ISBN 978-0-9893923-9-6. Available at: https://gifttutoring.org/documents/

Hu, L. T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: a Multidisciplinary Journal*, *6*(1), 1-55.

Hurst, E. J. (2015). Digital badges: Beyond learning incentives. *Journal of Electronic Resources in Medical Libraries*, *12*(3), 182-189.

Ifenthaler, D., Bellin-Mularski, N., & Mah, D. (Eds.). (2016). *Foundation of Digital Badges and Micro-Credentials: Demonstrating and Recognizing Knowledge and Competencies*. Switzerland: Springer International Publishing. doi:10.1007/978-3-319-15425-1

Itow, R. C., & Hickey, D. T. (2016). When Digital Badges Work: It's Not About the Badges, It's About Learning Ecosystems. In *Foundation of Digital Badges and Micro-Credentials* (pp. 411-419). Springer, Cham.

Jakobsson, M. (2011). The achievement machine: Understanding Xbox 360 achievements in gaming practices. *Game Studies*, *11*(1), 1-22.

Jakobsson, M., & Sotamaa, O. (2011). Special issue-game reward systems. *Game Studies*, *11*(1).

Jones, W. M., Hope, S., & Adams, B. (2017). Teachers' perceptions of digital badges as recognition of professional development. *British Journal of Educational Technology*, *49*(3), 427-438.

Jovanovic, J. & Devedzic, V. Tech Know Learn (2015) 20: 115. https://doi-org.pitt.idm.oclc.org/10.1007/s10758-014-9232-6

Kessels, U., Heyder, A., Latsch, M., & Hannover, B. (2014). How gender differences in academic engagement relate to students' gender identity. *Educational Research*, *56*(2), 220-229.

Kirschner, P. A., Sweller, J., & Clark, R. E. (2006). Why minimal guidance during instruction does not work: An analysis of the failure of constructivist, discovery, problem-based, experiential, and inquiry-based teaching. *Educational Psychologist*, *41*(2), 75-86.

Koedinger, K. R., Corbett, A. T., & Perfetti, C. (2012). The Knowledge-Learning-Instruction framework: Bridging the science-practice chasm to enhance robust student learning. *Cognitive Science*, *36*(5), 757-798.

Lave, J., & Wenger, E. (1991). *Situated Learning: Legitimate Peripheral Participation*. Cambridge, UK: Cambridge University Press.

Lee, C. D. (2008). Cultural Modeling as opportunity to learn: Making problem solving explicit in culturally robust classrooms and implications for assessment. In P. Moss, D. Pullin, J. P. Gee, E. Haertel, & L. J. Young (Eds.), *Assessment, Equity, and Opportunity to Learn* (pp. 136-169). NY: Cambridge University Press.

Lowendahl, J. (2015, July 8). Hype Cycle for Education, 2015. Retrieved September 7, 2015, from https://www.gartner.com/doc/3090218/hype-cycle-education-

LRNG. (2017) LRNG. Retrieved November 08, 2017, from https://www.lrng.org/about

LRNG. (n.d.). Products: LRNG. Retrieved September 4, 2016, from http://about.lrng.org/products/

LRNG. (n.d.). Products: LRNG. Retrieved September 4, 2016, from http://about.lrng.org/products/

MacArthur Foundation. (2011, September 15). Digital Media & Learning Competition Provides $2 Million for Innovations in Digital Badges. Retrieved September 7, 2015, from https://www.macfound.org/press/press-releases/digital-media-learning-competition-provides-2-million-for-innovations-in-digital-badges/

MacArthur Foundation. (n.d.). Connected Learning Principles. Retrieved September 3, 2015, from http://connectedlearning.tv/connected-learning-principles

MacArthur Foundation. (n.d.). Digital Badges. Retrieved September 3, 2015, from https://www.macfound.org/programs/digital-badges/

Malan, S. P. T. (2000). The 'new paradigm' of outcomes-based education in perspective. *Journal of Family Ecology and Consumer Sciences = Tydskrif Vir Gesinsekologie En Verbruikerswetenskappe*, *28*, 22-28.

Maslowsky, J., Jager, J., & Hemken, D. (2015). Estimating and interpreting latent variable interactions: A tutorial for applying the latent moderated structural equations method. *International Journal of Behavioral Development*, *39*(1), 87-96.

McDaniel, R., & Fanfarelli, J. (2016). Building better digital badges: Pairing completion logic with psychological factors. *Simulation & Gaming*, *47*(1), 73-102.

Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher*, *32*(2), 13–23.

Meyer, J. W., & Rowan, B. (1977). Institutionalized organizations: Formal structure as myth and ceremony. *American Journal of Sociology*, *83*(2), 340-363.

Meyer, J. W., & Rowan, B. (1978). The Structure of Educational Organizations. In Marshall W Meyer and Associates, *Environments and Organizations* (pp. 78-109). Jossey-Bass.

Mislevy, R. J. (2006). Issues of Structure and Issues of Scale in Assessment from a Situative/Sociocultural Perspective. CSE Technical Report 668. *National Center for Research on Evaluation, Standards, and Student Testing (CRESST).*

Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2003). Focus article: On the structure of educational assessments. *Measurement: Interdisciplinary Research and Perspectives, 1*(1), 3-62.

Moore, A. M., & Edwards, L. (2016). College and career ready. In Muilienburg, L.Y. & Berge, Z. L. (Eds.), *Digital Badges in Education: Trends, Issues, and Cases*, 111.

Moroder, K. (2014, April 7). Micro-Credentials: Empowering Lifelong Learners. Retrieved September 7, 2015, from http://www.edutopia.org/blog/micro-credentials-empowering-lifelong-learners-krista-moroder

Morrison, B. B., & DiSalvo, B. (2014, March). Khan academy gamifies computer science. In *Proceedings of the 45th ACM technical symposium on Computer science education* (pp. 39-44). ACM.

Mozilla Foundation. (2015). Issue | Open Badges. Retrieved September 7, 2015, from http://openbadges.org/issue/

Muthén, L. K. and Muthén, B.O. (1998-2017). Mplus user's guide. Eighth Edition. Los Angeles, CA: Muthén & Muthén.

O'Byrne, W. I., Schenke, K., Willis III, J. E., & Hickey, D. T. (2015). Digital badges recognizing, assessing, and motivating learners in and out of school contexts. *Journal of Adolescent & Adult Literacy*, *6*(58), 451-454.

Open Badges Project. (2017) Developers Guide. Retrieved November 08, 2017, from https://openbadges.org/developers/

Ososky, S. (2015). Opportunities and Risks for Game-Inspired Design of Adaptive Instructional Systems. In D. D. Schmorrow & M. C. Fidopiastis (Eds.), *Foundations of Augmented Cognition: 9th International Conference, AC 2015, Held as Part of HCI International 2015, Los Angeles, CA, USA, August 2-7, 2015, Proceedings* (pp. 640-651): Springer International Publishing.

Ostashewski, N., & Reid, D. (2015). A History and Frameworks of Digital Badges in Education. In *Gamification in Education and Business* (pp. 187-200). Springer International Publishing.

Peck, K. (2013). Reinventing the Report Card. Retrieved September 7, 2015, from http://www.advanc-ed.org/source/reinventing-report-card

Pintrich, P. R., & De Groot, E. V. (1990). Motivational and self-regulated learning components of classroom academic performance. *Journal of Educational Psychology*, *82*(1), 33.

Popper, K. (2005). *The logic of scientific discovery*. Routledge.

Pursel, B. K., Stubbs, C., Woong Choi, G., & Tietjen, P. (2016). Digital badges, learning at scale, and big data. In L. Y. Muilenburg & Z. L. Berge (Eds.), *Digital badges in education: Trends, issues, and cases* (pp. 93-101). Taylor & Francis.

Reid, A. J., Paster, D., & Abramovich, S. (2015). Digital badges in undergraduate composition courses: effects on intrinsic motivation. *Journal of Computers in Education*, *2*(4), 377-398.

Resnick, M. (2012). Still a Badge Skeptic. Retrieved September 7, 2015, from https://www.hastac.org/blogs/mres/2012/02/27/still-badge-skeptic

Rughinis, R. (2013, April). Talkative objects in need of interpretation. Re-thinking digital badges in education. In *CHI'13 extended abstracts on human factors in computing systems* (pp. 2099-2108). ACM.

Ryan, R. M., & Deci, E. L. (2000). Intrinsic and extrinsic motivations: Classic definitions and new directions. *Contemporary Educational Psychology*, *25*(1), 54-67.

Ryan, R. M., & Deci, E. L. (2000). Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. *American Psychologist*, *55*(1), 68.

Ryan, R. M., Mims, V., & Koestner, R. (1983). Relation of reward contingency and interpersonal context to intrinsic motivation: A review and test using cognitive evaluation theory. *Journal of Personality and Social Psychology*, *45*(4), 736.

Schiefele, U. (2009). Situational and individual interest. *Handbook of Motivation at School*, 197-222.

Schunn, C. D., McGregor, M. U., & Saner, L. D. (2005). Expertise in ill-defined problem-solving domains as effective strategy use. *Memory & Cognition*, *33*(8), 1377-1387.

Science Learning Activation Lab (2016a). *Measures Technical Brief: Engagement in Science Learning Activities (version 3.2).* Retrieved November 06, 2017, from http://www.activationlab.org/wp-content/uploads/2016/08/Engagement-Report-3.2-20160803.pdf

Science Learning Activation Lab (2016b). Measures Technical Brief: Fascination in Science (version 3.2). Retrieved November 07, 2017, from http://www.activationlab.org/wp-content/uploads/2016/03/Fascination-Report-3.2-20160331.pdf

Sternberg, R. J., & Horvath, J. A. (Eds.). (1999). *Tacit knowledge in professional practice: Researcher and practitioner perspectives*. Psychology Press.

Suhr, H. C. (2014). *Evaluation and credentialing in digital music communities: Benefits and challenges for learning and assessment*. MIT Press.

Torrance, H. (2012). Formative assessment at the crossroads: Conformative, deformative and transformative assessment. *Oxford Review of Education*, *38*(3), 323-342.

Tyack, D., & Tobin, W. (1994). The "grammar" of schooling: Why has it been so hard to change?. *American Educational Research Journal*, *31*(3), 453-479.

University of Minnesota. (2017) Open Microcredentials | Hype Cycle for Education. Retrieved January 15, 2017, from http://hypecycle.umn.edu/hype-cycle-technologies/open-microcredentials

University of Pittsburgh. *City of Pittsburgh Neighborhood Profiles American Community Survey Five-Year Estimates Update for 2006-2010 Data* (Rep.). (2012).

Van Buuren, S., Groothuis-Oudshoorn, K. (2011). mice: Multivariate Imputation by Chained Equations in R. *Journal of Statistical Software*, *45*(3), 1-67. http://www.jstatsoft.org/v45/i03/

Wardrip, P.S., Abramovich, S., Bathgate, M. & Kim, Y. J. (2016). A school-based badging system and interest-based Learning: An exploratory case study. *International Journal of Learning and Media*. http://www.acsu.buffalo.edu/~samuelab/IJLM_Badge_Paper.pdf

Whitmore, P. G., & Fry, J. P. (1974). *Soft skills: Definition, behavioral model analysis, training procedures* (No. HumRRO-PP-3-74). HUMAN RESOURCES RESEARCH ORGANIZATION ALEXANDRIA VA.

Wigfield, A., & Eccles, J. S. (2000). Expectancy–value theory of achievement motivation. *Contemporary Educational Psychology*, *25*(1), 68-81.

Wiggins, G. (1998). *Educative Assessment. Designing Assessments To Inform and Improve Student Performance*. San Francisco, CA: Jossey-Bass Publishers.

Wiggins, G., & McTighe, J. (2011). What is backward design? *Understanding by Design*, 7-19.

Witherspoon, E. B., Schunn, C. D., Higashi, R. M., & Shoop, R. (2018). Attending to structural programming features predicts differences in learning and motivation. *Journal of Computer Assisted Learning*, *34*(2), 115-128.

Wood, D., Bruner, J. S., & Ross, G. (1976). The role of tutoring in problem solving. *Journal of Child Psychology and Psychiatry*, *17*(2), 89-100.

Yang, J. C., Quadir, B., & Chen, N. S. (2016). Effects of the badge mechanism on self-efficacy and learning performance in a game-based English learning environment. *Journal of Educational Computing Research*, *54*(3), 371-394.