Perceived Relevance of Digital Badges Predicts Longitudinal Change in Program Engagement

Higashi, R. M., & Schunn, C. D. (In press). Perceived relevance of digital badges predicts longitudinal change in program engagement. *Journal of Educational Psychology*

ABSTRACT

Digital badges have long been assumed to possess motivational qualities that could encourage learners to engage with learning content. However, prior studies have found the effects of badges to be complex, differing by learner, type of badge, and potentially other factors. Qualitative reports suggest that individuals' perceptions of digital badges may play a role in moderating badges' effects: badges are only motivating when they are perceived as relevant and desirable. In the current study, we examine longitudinal episodes of data from 2,410 middle and high school users of a badged online programming curriculum to test whether there is evidence that learners' perceptions of badges' relevance predict changes in engagement over time; and whether that relationship is equitable with respect to age, sex, and ethnicity. We also investigate whether a reciprocal relationship may exist in which engagement predicts relative increases in learners' perceptions of badges as relevant to them. Learners' positive perceptions of badges' relevance predicted rank-order increases in engagement over time. Further, this relationship did not vary by age, sex, or minoritized racial/ethnic status. In addition, higher engagement also predicted upward shifts in perceived badge relevance. These results suggest that learners' subjective evaluations of digital badges are closely related to changes in their engagement with program activities, that badged engagement neither widens nor closes educational equity gaps, and that learners' regard for badges and engagement in program activities may be mutually reinforcing.

Keywords: Digital badges; Engagement; Perceived relevance

Educational Impact and Implications Statement

Digital badges have been making inroads into classrooms and out-of-school time programs for nearly a decade, but their real-world effectiveness as motivational tools vary widely. In this study, we modeled individuals' opinions of badges, and found that learners who perceive a program's badges as sensible,

valuable, and desirable become increasingly engaged with that program activities over time. This effect was equally true across age, sex, or race/ethnicity subgroups. Learners who are more engaged also tended to find the badges increasingly relevant over time, creating a “virtuous cycle” of increasing program engagement and perceived badge relevance over the course of the program. Thus, educational programs may be able to use digital badges to sustain and deepen student engagement, provided they create badges that participants find sensible, valuable, and desirable.

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 racial/ethnic

background; 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.

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-providers 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).

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. Our work extends this trajectory to begin unpacking the "black box" previous analyses have left around the *individual* by connecting internal factors to external ones.

Perceived Badge Relevance

Prior qualitative work points to *participant perceptions of badges' relevance* as a pivotal aspect of this connection. Davis and Singh (2015) report on a school-based program whose participants judged its non-credit-earning badges to be pointless, and so ignored them in favor of activities that would yield better grades. The badges, perceived as lacking value, had no effect on students' activity. Wardrip, Abramovich, Bathgate, and Kim (2016) interviewed a pilot cohort of students in a school-based badge system, noting that students' appraisal of the badges' meaning and value were connected to their assessments of the badges not being trivial to earn. Suhr (2014) likewise noted some members of an online music-making community dismissed its badges as undesirable because they were earned for achievements they considered trivial. In each of these cases, it is clear that potential badge earners are making value judgments about the badge systems, which then predicate their engagement with the badges. Learners with more positive subjective evaluations of a program's badges are more likely to engage with them, and those subjective evaluations appear to center around a perception of the badges as "relevant" in terms of desirability, value, and meaningfulness in the program context.

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. 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 this initial 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.

Theory of action

We theorize that a positive subjective evaluation of a program's badges sets off a theoretical "chain reaction". Learners who regard a program's badges as more relevant (high PBR) become more engaged with the badges. This causes them to evince more of the badges' effects. As most badge systems are designed to increase learner engagement in a program's activities, this means learners with more positive subjective evaluations of the badging system will engage more with (engagement-increasing) badges. This results in 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]

PBR and engagement. The term "engagement" has been used to refer to many different constructs. 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. Many studies, especially of large-scale online courses (i.e. MOOCs), have exclusively focused on this kind of engagement by using participation metrics (e.g., number of logons, amount of time in the system) as measures of engagement. *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).

There is some preliminary evidence of a positive relationship between PBR and Engagement. Higashi (2018) 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.

PBR and domain motivation. Motivation is theoretically and practically distinct from engagement – in fact, simply put, engagement is the *consequence* of motivation. Domain motivation captures learner orientations toward a particular learning domain, e.g. science or math. We argue that PBR is conceptually distinct from domain motivation because it is specifically *about badges*. But there remains an empirical and theoretical possibility that any relationship between PBR and engagement could simply be a side effect of other, well known motivational factors such as domain interest. Under this opposing hypothesis, PBR and engagement are simple, otherwise unrelated consequences of a student being interested in the learning domain. If such were the case, empirical evidence would show the relationship between PBR and Engagement being explainable – in part or whole – by these other motivational factors. In a correlational analysis, including the motivational factor would "explain away" any effect of PBR on engagement. Higashi (2018) previously found that perceived success during program activities, another motivational construct, was related to PBR, but each predicted distinct variance in engagement. However, perceived success is only a short-term motivational factor; persistent long-term motivational covariates such as interest or identity, each of which is a known predictor of

engagement, are more likely candidates to supercede PBR as an originating effect. We describe these in detail below.

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.

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 and ethnic 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 racial/ethnic background, sex, and age.

Positive feedback loops. Researchers have theorized that recursive feedback mechanisms may be 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.

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 racial/ethnic background. 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.

Methods

We investigated our research questions using three samples drawn from a common data set. We first describe overall sample characteristics, the learning context in which the data was collected, the measures contained in the data, and the longitudinal unit of analysis we synthesized from the data. We

will then describe the analyses used to answer each question individually, along with the sample used in each.

Participants

Our sample includes all users of the online Computer Science STEM Network (CS2N) curriculum platform who 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. Because demographic data on the CS2N service are self-reported and optional, we cannot be sure of the exact composition of users in the sample. However, extrapolating from prior teacher surveys and analyses of server logs, participants are generally located across the US, and most frequently come from middle school technology, robotics, and computer science classes, but sometimes also from early high school technology classes. Self-reported account registration data on the server corroborate this, with a median participant age of 14 years, which is consistent with a mixture of middle and high school students (although grade level was not recorded). Students in the main sample self-reported as 64% male, 32% female, and 4% blank or undisclosed. This 2:1 ratio of male to female students is in the same range as CS2N's overall user base, which has a 2.5:1 self-reported male-to-female ratio. The Group membership constraint means that our sample primarily consists of students in classrooms or organized clubs rather than independent settings, although the range of group sizes (min = 1, max = 125) suggests that usage contexts still included some home users and small afterschool programs, as well as larger classroom and multi-section groupings. We further narrowed this sample for individual analyses based on the data required for each, and describe them separately in the Procedures section.

Data Collection

Context. 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 2 shows the simulator interface, the coding interface, and the curriculum interface as seen by the user.

[Figure 2]

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.

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.

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 et al., 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”.

Collection. 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.

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 3. 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).

[Figure 3]

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/ethnicity. 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.

Measures

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 *Procedure* 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 3 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 ($\alpha = .72$). 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.

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, 2018).

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.

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.

Demographic data. Participants who completed the pre-survey (see *Procedure* below) reported their age (in years), sex (Boy or Girl), and racial/ethnic background (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/ethnicity response; these 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).

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.

Research Design

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 4. As it comprises a temporal sequence of Engagement-Badges-Engagement measures, we refer to it as an EBE Tri.

[Figure 4]

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.

Note that the longitudinal ("time") unit in the data is not *days of class time*, but *lessons of curricular progress*. Class schedules and individual rates of progress vary, but data are collected at fixed

points of student curricular progress. Accordingly, the interpretation of relationships within a Tri are episodic: they are relationships between prior engagement, PBR, and subsequent engagement for a particular slice of content. When pooling data from Tris at different points in the curriculum, we make the assumption that the relationships between PBR, prior engagement, and subsequent engagement are substantively similar across episodes drawn from different parts of the curriculum. We also tested this pooling assumption in each analysis by verifying that the patterns replicated within each of the particular curriculum Tris.

Tris in the CS2N Curriculum. We were able to rule out one additional threat to this assumption, that Engagement or PBR data from adjacent data points might not be comparable due to innate, idiosyncratic differences in lesson 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).

One final 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. Learners in our study covered an average of 5.2 lessons, with a median of 4 lessons, out of a potential 14. While the lessons covered were typically consecutive, it was not always the same consecutive set. Approximately one-third of learners completed only a single lesson. This form of heterogeneity is typical instructional behavior in 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 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 5.

[Figure 5]

As a final note, we included partial Tris in the reported analyses because even a Tri that is missing one of its three points contributes usable information in estimating the correlation of the remaining two. Full Information Maximum Likelihood estimation does so automatically. We also verified whenever possible that results were substantively similar under listwise deletion.

Analysis 1a (Longitudinality)

Sample. This analysis uses all available EBE Tris in the data, resulting in a sample of n=3,696 partial and complete Tris from 2,410 users in 189 groups. EBE Tris came from four segments of the curriculum (see Appendix B). Specific demographic characteristics of this sample were not available because most users did not include them in their online profiles.

Procedure. 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. Note that separate models of each Tri within the curriculum produced similar results as the analyses which pooled all the Tris together, although with larger standard errors due to the small number of data points.

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. Ultimately, all models produced similar results, so for parsimony, we report the single-level SEM result as the final model, and provide details of the others in Appendix C. All analyses were performed using Mplus Version 8 and maximum likelihood estimation with robust standard errors (Muthen & Muthen, 2017).

Analysis 1b (Demographics and Motivation)

Sample. 10 teachers were recruited from a summer teacher professional development program for the curriculum. These teachers distributed paper pre-surveys that included demographic questions to their students. 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.

The surveys indicate that this subsample represents a predominantly middle-school population – 49% female, with a mean age of 12.6 years (SD=1.0) – organized into classroom-sized units with a median size of 19 students (mean = 29 students). 12% of subsample respondents reported racial backgrounds that are underrepresented in technology fields (non-White or Asian).

Procedure. To refine our model and test whether motivational confounds may be responsible for observed relationships, we analyze the subset of data for which motivational pre-survey data is available. We begin by verifying that the final model from the previous analysis is 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.

Analysis 2 (Equity)

Sample. This analysis uses the same subsample of pre-surveyed learners as Analysis 1b (1,832 EBE Tris from 458 users in 38 groups).

Procedure. 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 racial/ethnic 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. Where results indicate that the models are not substantively different, we also verify that no radical changes in other (e.g. main effect) estimates occurred. 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 6, using dotted lines to identify the portion that is different between the models.

[Figure 6]

Analysis 3 (Reciprocality)

Sample. This analysis uses the sample of "BEB Tris" contained in the original dataset: *n*=6,838 BEB Tris from 2,980 users in 236 groups. BEB Tris came from 5 locations in the curriculum. As with the full EBE Tri sample, demographic information is unavailable for this BEB Tri sample.

Procedure. Our final research question is whether we see evidence of reciprocal effects in which higher engagement predicts a relative rise in Perceived Badge Relevance. We apply the same model-building approach as in Analysis 1a to the set of "BEB" Tris to examine whether higher levels of engagement predict relative change in perceptions of badge relevance.

Results

In this section, we present the substantive highlights of our findings; for a detailed description of additional models and integrity checks, please see Higashi (2018).

Analysis 1a

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 1]

SEM Model. Our final model is shown in Figure 7. 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 7]

Analysis 1b and Analysis 2

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 2.

[Table 2]

Compared to the full sample, the sub-sample had higher mean Engagement and PBR, but the magnitude of these differences was negligible (see Appendix D). Most importantly, correlations between the Engagement and PBR factors resemble those in the full sample, suggesting the two samples were functionally equivalent in terms of the phenomenon of interest. Motivational items were highly correlated with each other and moderately correlated with Engagement, but only slightly correlated with PBR. Age has 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 moderately high correlations (around .6) with each others' items, suggesting that there would be some danger of multicollinearity when including both constructs in the same model.

Motivational Predictors (Analysis 1b). We began with the final model from the previous analysis (i.e. with no motivational predictors), which had acceptable fit in the pre-survey matched subset of the data (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 $\beta = .22$, with pseudo-R^2 = 65% of $Engagement_{t+1}$ variance explained.

Adding interest, identity, age, and minoritized racial/ethnic status to the model (n=1,832) initially produced only marginal fit due to a high correlation between Identity and Interest (r = .72), indicating a likely multicollinearity issue. Identity was removed, resulting in the model shown in Figure 8, which 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. We therefore describe only the Interest-based model moving forward.

[Figure 8]

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 slightly less common overall (r = -.02, 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 = -.06, p = .08). Perceived Badge Relevance, PBR_t, was predicted by prior-timepoint engagement (β = .59, p < .001), but only marginally by sex (β =.07, p = .06) and racial minoritization status (β =.06, 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 = -.05, 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.

Equity Interactions (Analysis 2). To test whether the impact of digital badges appears to be equitable by sex, age, and race/ethnicity, 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 3.

[Table 3]

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. Even in a worst-

plausible-case scenario where we assume that each interaction effect was at its two standard error bound, the differences would not have been enough to threaten the statistical significance of the main effect. 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 racial/ethnic status within our sample.

Analysis 3

Descriptive Statistics. A list of means, standard deviations, and bivariate correlations are shown in Table 4. 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 4]

SEM Model. We use a mirrored version of the final model from Analysis 1 to address the question of reciprocality directly. Our final model is shown in Figure 9. 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 higher observed engagement tended to predict relative increases in PBR over time. For example, if two learners had the same PBR prior to a lesson, but one reported an engagement score 1 point higher during the lesson, the more engaged learner would be predicted to have a subsequent-timepoint PBR score .22 points higher than the less-engaged one. The overall model explains pseudo-R^2 = 54% of variance in PBR_{t+1}.

[Figure 9]

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/ethnicity, sex, and age; and (RQ3) whether the relationship between PBR and engagement appears to be reciprocal.

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=.16 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, 2018). 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 racial/ethnic 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.

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/ethnicity, 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.

This is consistent with our broader theoretical model, in which higher PBR leads to more positive subjective evaluations, which lead to greater engagement with program badges, allowing those badges to be effective in their role of increasing program engagement. The finding of reciprocal reinforcement from Engagement back to PBR is also consistent with the cyclical model.

The positive direction of the PBR-to-Engagement relationship is also consistent with the findings of Abramovich, Schunn, and Higashi (2013). The type of badge design used in this study were "skill badges" that connected directly to quality of performance, as opposed to being "participation badges" which earned no matter what. Skill badges should behave as intrinsic rather than extrinsic motivators. Learners with higher PBR engaged more strongly with performance-contingent skill badges, leading them to engage more actively in the corresponding activities.

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/ethnic 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 ($\beta = .07$, $p=.06$) and racial/ethnic background ($\beta = .06$, $p=.11$). 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 ($\beta = .59$) 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 ($\beta = .41$), age ($\beta = -.14$), and marginally by minoritized racial/ethnic status ($\beta = -.06$, $p = .08$); 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/ethnic 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.

Does Engagement Also Predict Subsequent PBR?

While it's intuitive to understand that PBR reflects learner receptiveness to badge effects, it is not *prima facie* 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. Yet, not only is the relationship in this direction significant ($\beta = .24$, SE=.02), it is comparable in size to (or even slightly larger than) the predictive effect of PBR on engagement ($\beta = .16$, SE=.02). This result is 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. Taken together, our findings are 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.

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). While reflective of the real-world usage patterns of the CS2N system in which teachers pick and choose sections of curriculum to use with their students, the average user in our study contributed between one and two EBE Tris, corresponding to a few content units' worth of use. 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.

A related limitation is that, while our data comes directly from learner engagement with program activities, its resolution is limited to activity completions that triggered the earning of a badge. Engagement, of course, includes effort that does not quite pass the threshold of completion. Thus, future datasets should aim to provide higher-resolution measurements of learner engagement with program activities, including partial completion.

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.

APPENDIX A: Similarity of Engagement and PBR means across lessons

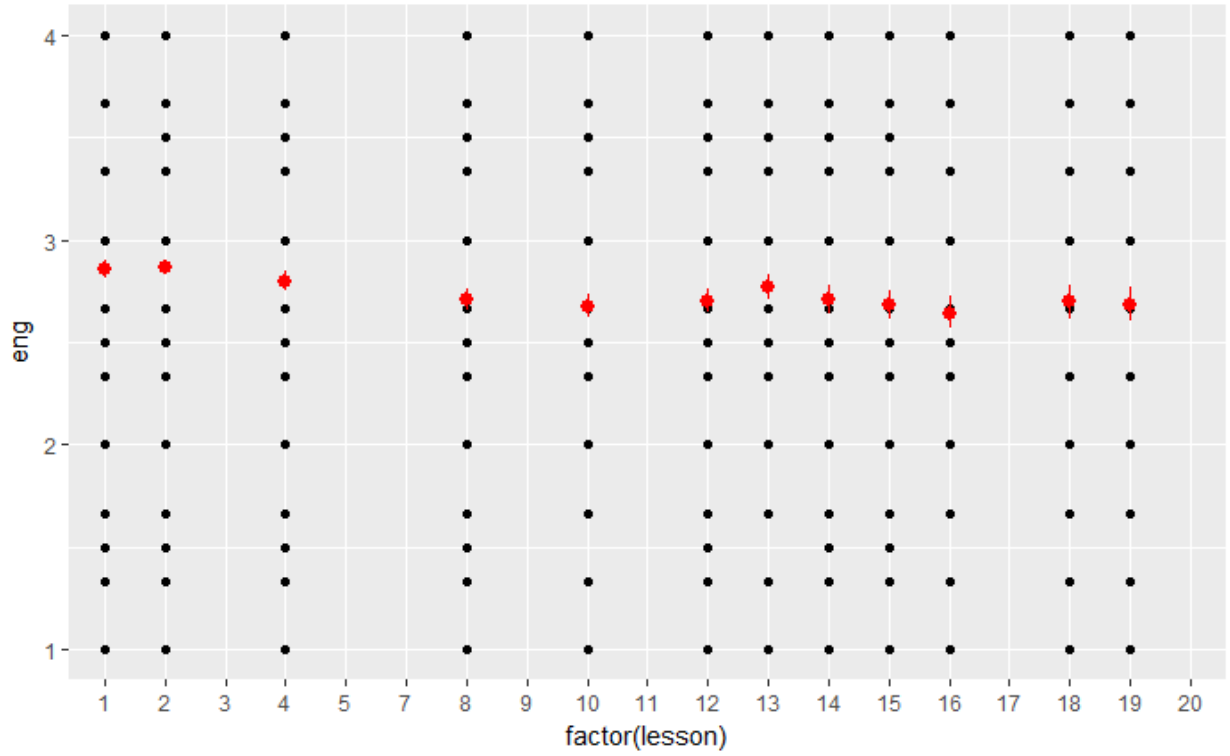

Figure A1. Plot of Engagement scores by lesson. Mean scores across Tris for each lesson are in red. Lines indicate bootstrapped 95% confidence intervals for means.

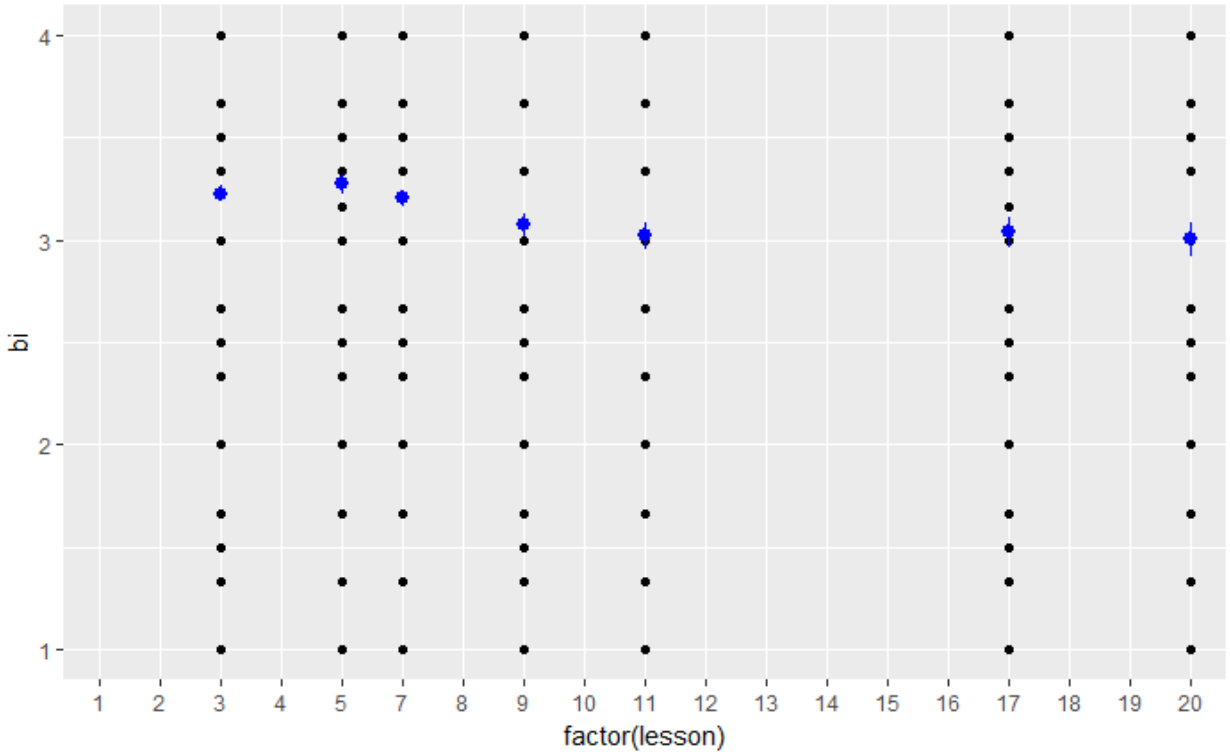

Figure A2. Plot of PBR (called "bi" in the dataset) scores by lesson. Mean scores across Tris for each lesson are in blue. Vertical lines indicate bootstrapped 95% confidence intervals for means.

APPENDIX B: 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 C: 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.

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 α = .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).

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 A1.

Table A1. *Intraclass correlations and covariance and correlation matrices for latent classroom-level intercepts in the two-level model.*

		Covrariance					Correlation				
	ICC	Eng_{t-1}	$PBR1_t$	$PBR2_t$	$PBR3_t$	Eng_{t+1}	Eng_{t-1}	$PBR1_t$	$PBR2_t$	$PBR3_t$	Eng_{t+1}
Eng_{t-1}	.146	0.09					-				
$PBR1_t$	.085	0.07	0.08				0.82	-			
$PBR2_t$	.116	0.09	0.10	0.13			0.85	0.943	-		
$PBR3_t$	.112	0.08	0.09	0.11	0.11		0.79	0.948	0.965	-	
Eng_{t+1}	.151	0.100	0.07	0.10	0.08	0.10	1.00	0.805	0.838	0.779	-

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 C1.

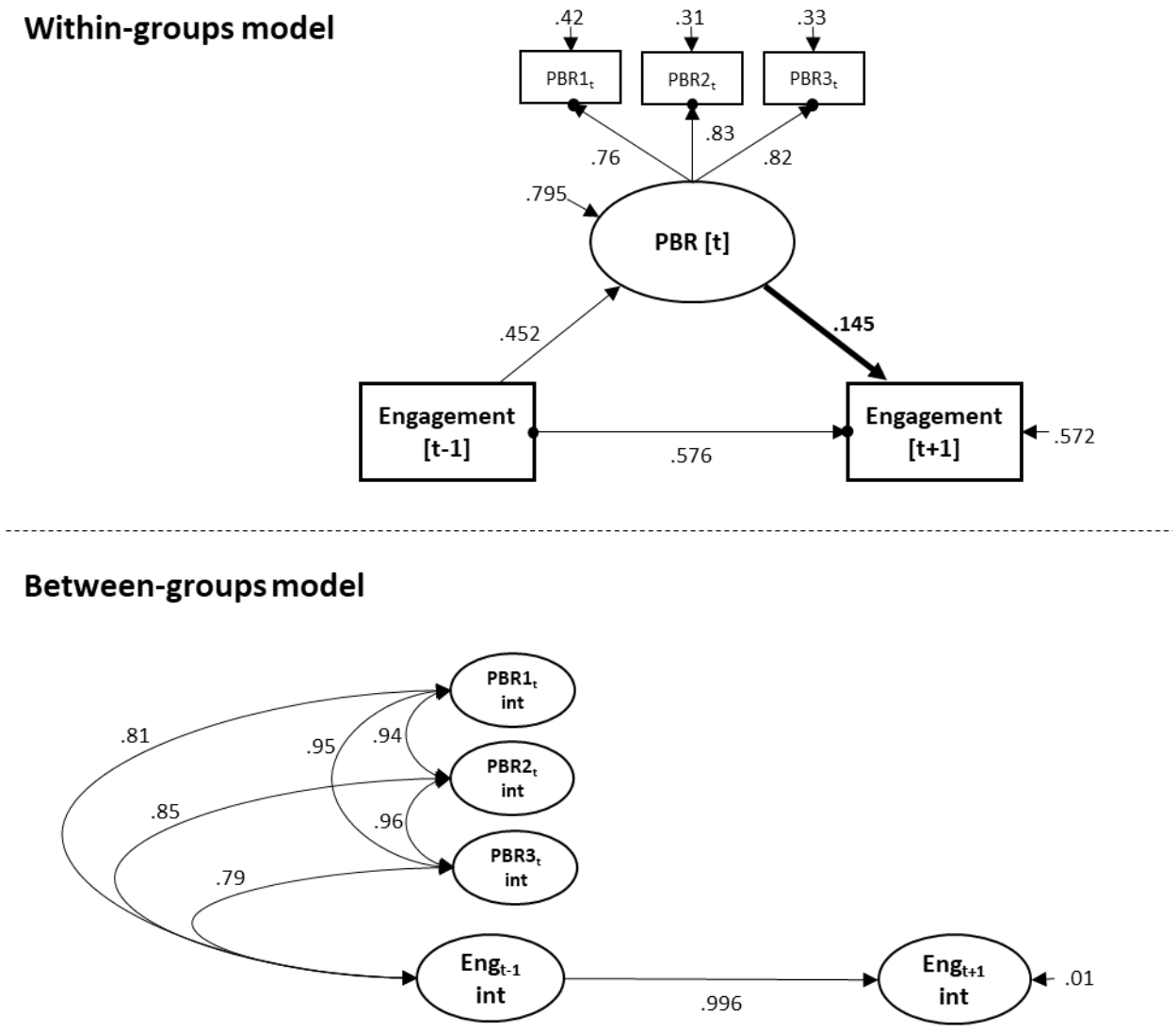

Figure C1. Two-level structural equation model.

The resulting model was substantively very similar to the single-level SEM model, estimating a standardized β=.14 for the effect of PBR_t on $Engagement_{t+1}$, and also similar estimates for other factors.

APPENDIX D. Full sample vs. surveyed subsample comparisons

The subsampling procedure used to select participants for the paper-survey analyses was exogenous, but not random. Unlike the online surveys, the paper pre-surveys were delivered to participants through their instructors, whom we recruited through a professional development session. Thus, we expected that the surveyed subsample will be biased in some respects compared to the population of CS2N users: they will be students in supervised K-12 classrooms, with trained instructors who were interested and well-resourced enough to attend a robotics PD session at a university campus. Therefore, it is important to test whether this subsample was similar to the overall sample on observed responses, as this could affect the interpretation of the results. Table D1 below compares the common data – prior timepoint engagement, mean PBR score, and subsequent time point engagement – between the full sample and the subgroup which received the survey. Given the very large sample sizes, we use Cohen's *d* to judge whether the differences between the subsample and the full sample are substantive. Cohen's *d* differences of less than 0.20 are generally considered negligible (Cohen, 1988).

Table D1. Factor means, standard deviations, and Cohen's d *estimates for differences between full sample and surveyed subsample of participants.*

	Full sample (n=3,696)	Survey subsample (n=1,832)	Cohen's *d*
Engagement_{t-1}	2.74 (SD=.80)	2.87 (SD=.86)	.16
PBR_{t} (mean of factors)	3.11 (SD=.88)	3.23 (SD=.86)	.13
Engagement_{t+1}	2.71 (SD=.84)	2.80 (SD=.90)	.11

REFERENCES

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*, 217-232.

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

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, engineer- ing, and medicine. Journal of Research in Science Teaching, 47(5), 564– 582. http://doi.org/10.1002/tea.20353.

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.

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.

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

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.

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.

Cohen, J. (1988). Statistical power analysis for the behavioral sciences. 2nd edition. Routledge.

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.

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 Educatino, 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.

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., 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.

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).

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.

Ellis L.E., Nunn S.G., Avella J.T. (2016) Digital Badges and Micro-credentials: Historical Overview, Motivational Aspects, Issues, and Challenges. In: Ifenthaler D., Bellin-Mularski N., Mah DK. (eds.) Foundation of Digital Badges and Micro-Credentials. 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.

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.

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.

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

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*.

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. (2018). *The effect of perceived relevance of digital badges on student engagement* (Doctoral dissertation, University of Pittsburgh).

Higashi, R., M., Schunn, C. D., Nguyen, V. H., & Ososky, S. J. (2017). Coordinating evidence across learning modules using digital badges. In R. A. Sottilare, A. Graesser, X. Hu, & G. Goodwin (Eds.), Design Recommendations for Intelligent Tutoring Systems: Volume 5 - Domain Modeling. Orlando, FL: U.S. Army Research Laboratory.

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.

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).

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.

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

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/

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.

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

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

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.

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.

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* , available at: 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). 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.

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 (2016c). 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

Suhr, H. C. (2014). *Evaluation and Credentialing in Digital Music Communities: Benefits and Challenges for Learning and Assessment*. MIT Press.

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.

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

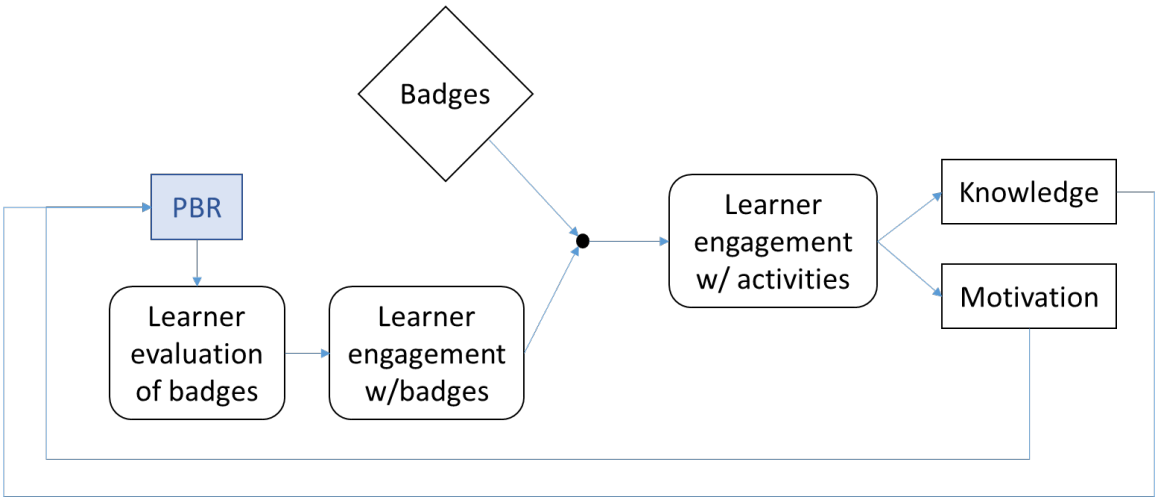

Figure 1. Theoretical model of action for learners' subjective evaluation of badges impacting program engagement and subsequent outcomes. Perceived Badge Relevance (PBR) is a predictor of learners' evaluations of badges.

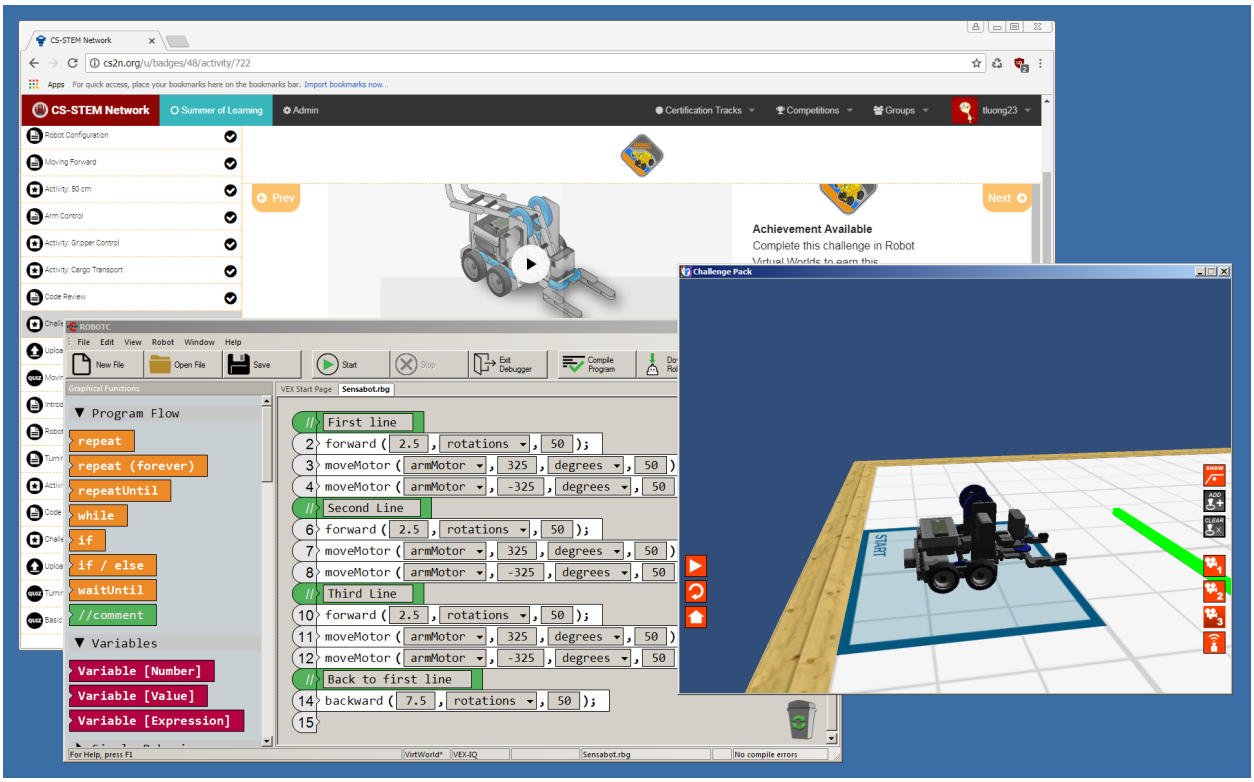

Figure 2. Software interface from the learner's perspective. Top: CS2N course materials; left bottom: ROBOTC programming software; right bottom: Robot Virtual Worlds 3D simulation environment.

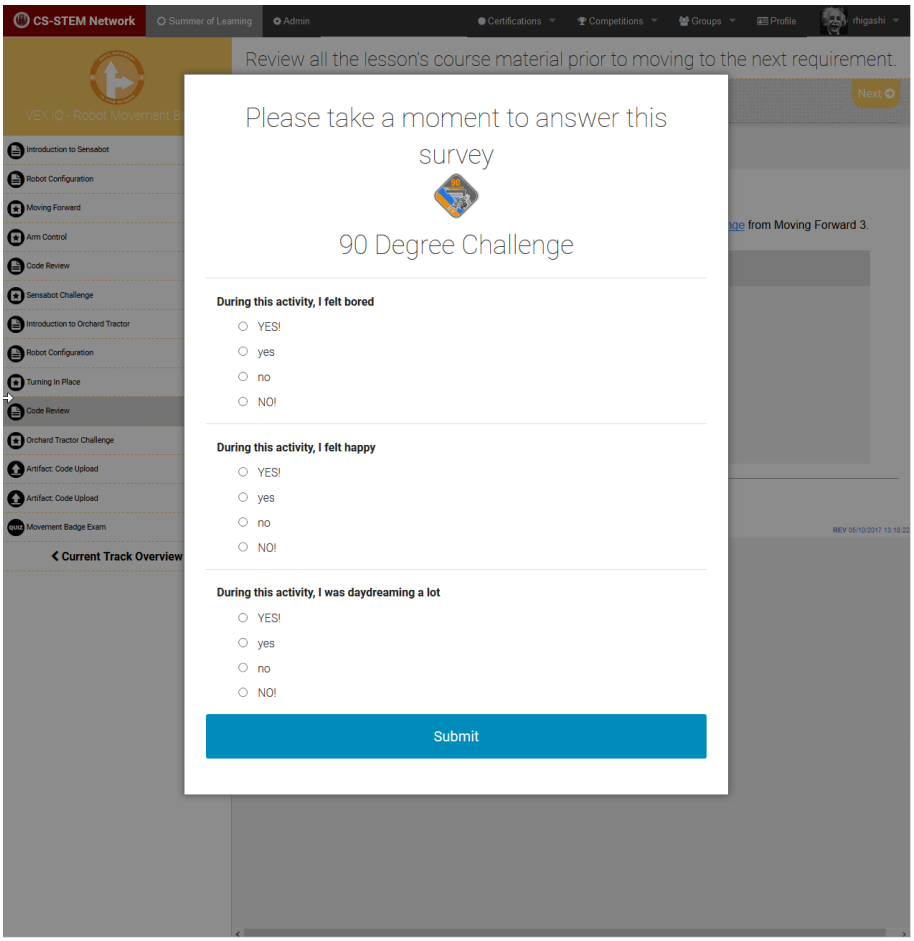

Figure 3. An Engagement survey as seen by student users. “90 Degree Challenge” is the name of both the Badge and the mini-challenge activity. The survey is overlaid onto the first CS2N page students view after completing the activity.

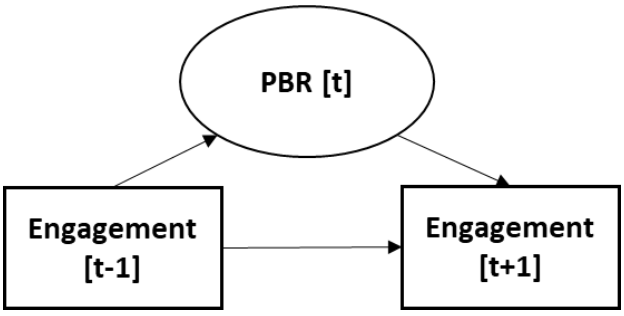

Figure 4. An Engagement-Badges-Engagement (EBE) Tri.

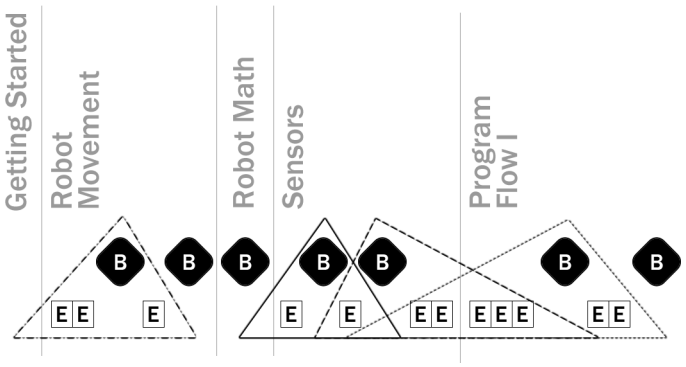

Figure 5. Data collection points across the curriculum. E = Engagement measured after a mini-challenge. B = Perceived Badge Relevance measured after a chapter challenge. Triangles represent EBE Tris; BEB Tris are not shown, but are simply the reciprocal pattern.

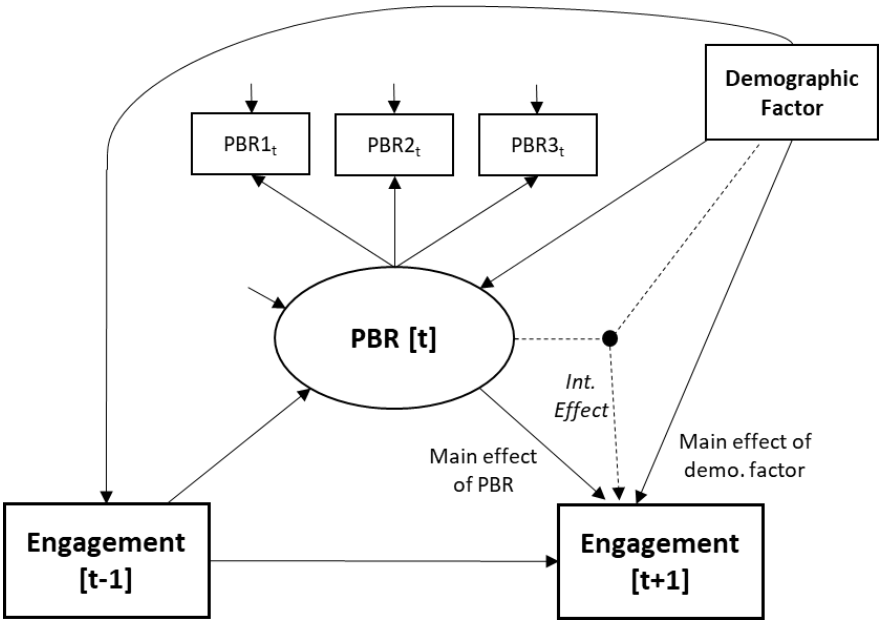

Figure 6. Diagram of the models used to test interaction between each demographic factor and the PBR effect on engagement. In the main effect model, the interaction effect (dotted line portion) is not present.

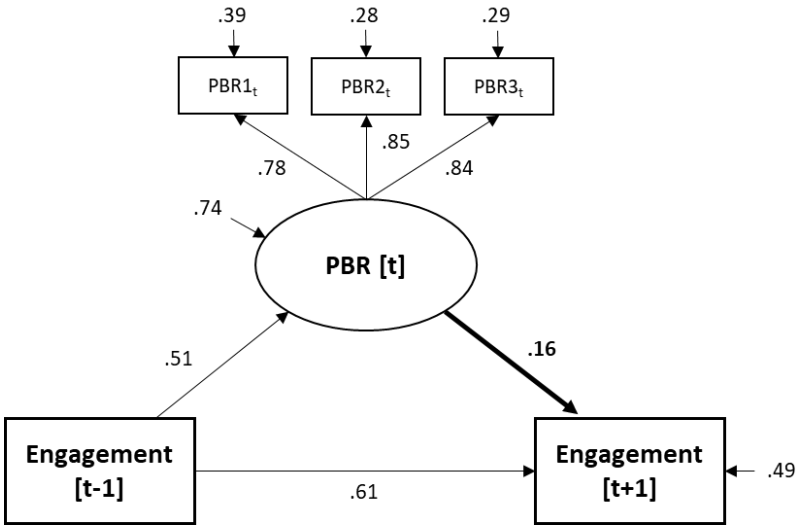

Figure 7. Diagram of final model. Path weights indicate standardized βs. All values are significant at $p <$ .001.

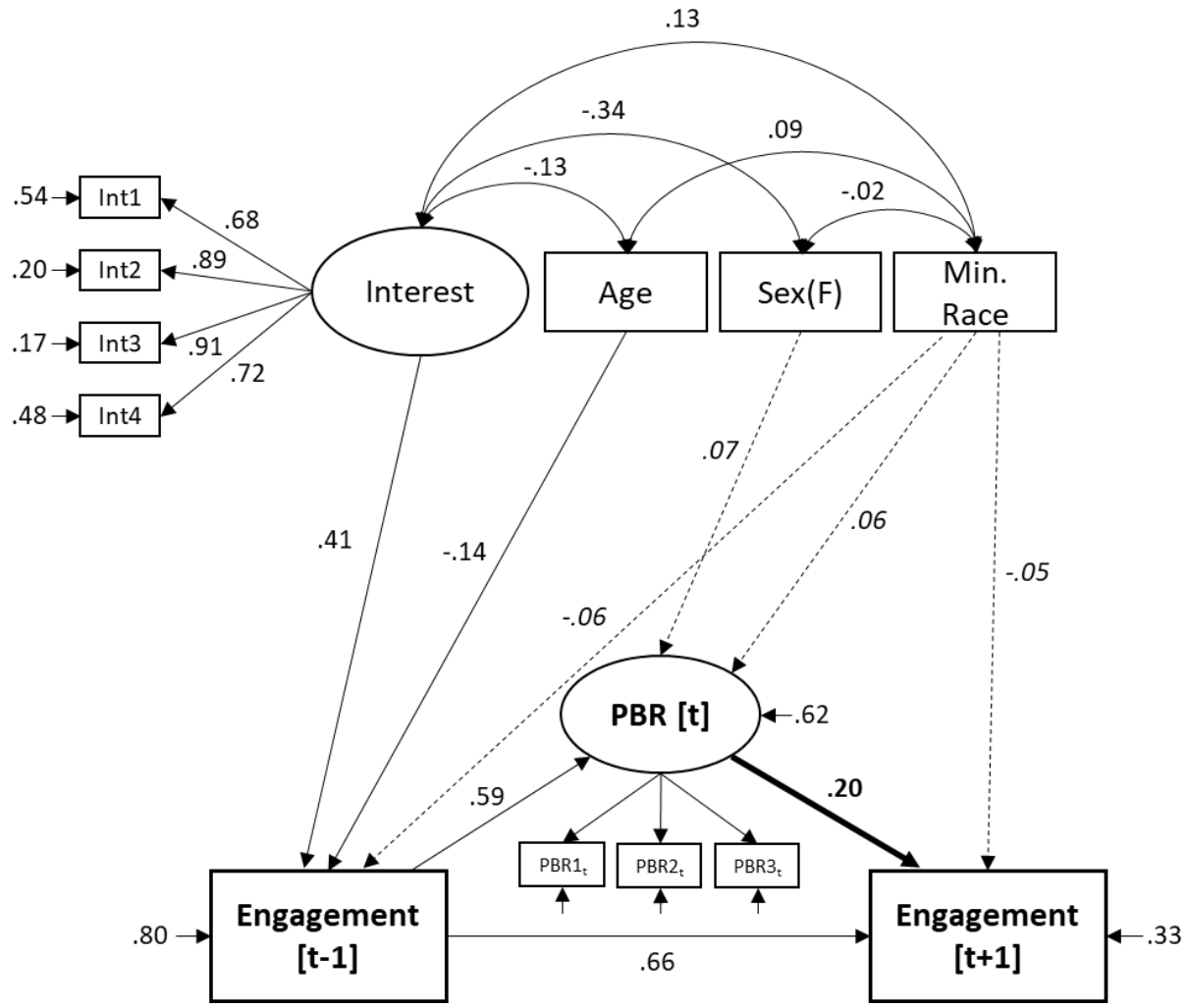

Figure 8. Model diagram for the final demographic model. All demographic and motivational predictors were estimated as predictors of $Engagement_{t-1}$, PBR_t, and $Engagement_{t+1}$, but only significant links are shown. Dotted lines represent marginally significant relationships. Coefficients represent standardized effects of continuous predictors, or contrast effects of categorical predictors in standardized units of the predicted values.

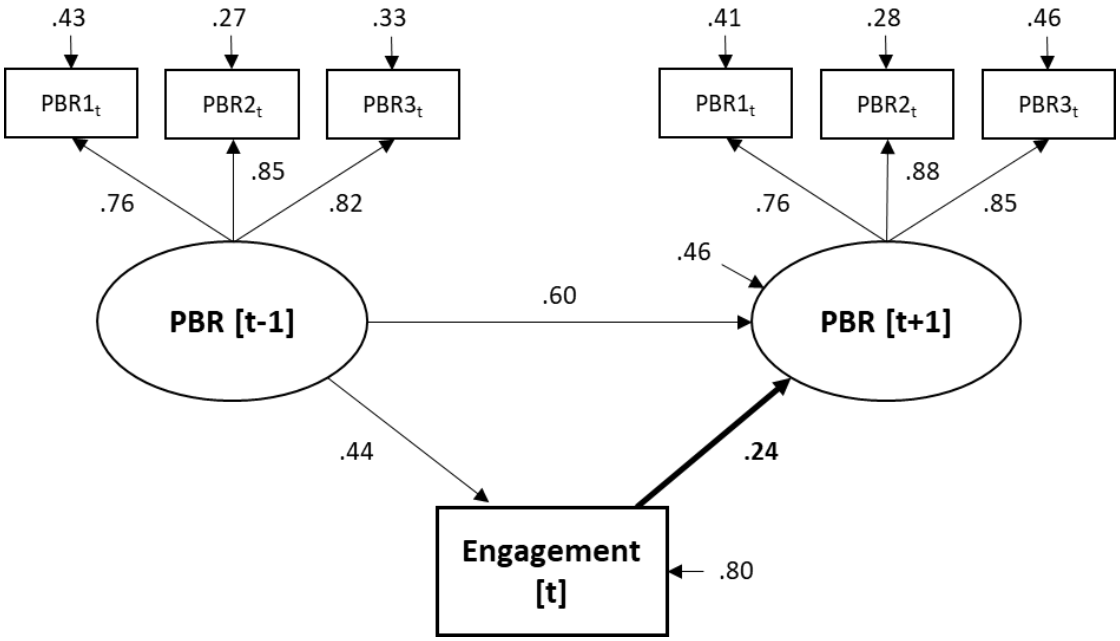

Figure 9. Diagram of final reciprocal (BEB) model. Path weights indicate standardized βs. All values are significant at $p < .001$.

Table 1. *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

Table 2. *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

Note. Multi-item scales are collapsed to their means here, but were modeled separately using latent factors. See Witherspoon et al. (2018) for item-level detail.

Table 3. *Model fit and unstandardized regression coefficients for the effect of sex, age, minoritized racial/ethnic status, and the interactions of those factors with PBR_t on $Engagerment_{t+1}$, controlling for $Engagement_{t-1}$ and PBR_t.*

	Sex (female)		Age (years)		Minoritized racial status	
	Main only	Interaction	Main only	Interaction	Main only	Interaction
AIC	7180.6	7182.4	7636.2	7637.4	7505.6	7507.5
SSA-BIC	7211.8	7215.2	7668.3	7671.2	7537.4	7541.0
Log-likelihood	-3571.3	-3571.2	-3799.1	-.3798.7	-3733.8	-3734.7
[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*

*Note. ns p>=.10, ~ p<.10, * p<.05, ** p<.01, *** p<.001.*

Table 4. *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