Int J Artif Intell Educ (2019) 29:58–88
https://doi.org/10.1007/s40593-018-0168-1

ARTICLE

Personalizing Algebra to Students' Individual Interests in an Intelligent Tutoring System: Moderators of Impact

Candace Walkington[1] · Matthew L. Bernacki[2]

Published online: 22 August 2018

Abstract Students experience mathematics in their day-to-day lives as they pursue their individual interests in areas like sports or video games. The present study explores how connecting to students' individual interests can be used to personalize learning using an Intelligent Tutoring System (ITS) for algebra. We examine the idea that the effects of personalization may be moderated by students' depth of quantitative engagement with their out-of-school interests. We also examine whether math problems designed to draw upon students' knowledge of their individual interests at a deep level (i.e., actual quantitative experiences) or surface level (i.e., superficial changes to problem topic) have differential effects. Results suggest that connecting math instruction to students' out-of-school interests can be beneficial for learning in an ITS and reduces gaming the system. However, benefits may only be realized when students' degree of quantitative engagement with their out-of-school interests matches the depth at which the personalized problems are written. Students whose quantitative engagement with their interests is minimal may benefit most when problems draw upon superficial aspects of their interest areas. Students who report significant quantitative engagement with their interests may benefit most when individual interests are deeply incorporated into the quantitative structure of math problems. We also find that problems with deeper personalization may spur positive affective states and ward off negative ones for all students. Findings suggest depth is a critical feature of personalized learning with implications for theory and AI instructional design.

Keywords Personalization · Personalized learning · Individual interest · Intelligent tutoring systems

✉ Candace Walkington
cwalkington@smu.edu

Matthew L. Bernacki
matt.bernacki@unlv.edu

[1] Southern Methodist University, 3101 University Blvd., Ste. 345, Dallas, TX 75025, USA

[2] University of Nevada, Las Vegas, 4505 S. Maryland Parkway, Las Vegas, NV 89154-3003, USA

Introduction

Students' interest in learning mathematics can decline as they reach adolescence (Fredricks and Eccles 2002; Frenzel et al. 2010), as the mathematics they are learning in school becomes less concrete and immediately relatable to everyday activity. Middle and high school mathematics can alienate students, which is particularly problematic due to its central importance for college and career readiness. Specifically, Algebra I has been identified as a "gatekeeper" to higher level mathematics courses, and to many careers (Kaput 2000; Moses and Cobb 2001; National Mathematics Advisory Panel 2008). However, students' confidence in using mathematics and their perception that math is useful both decline during algebra courses (McCoy 2005).

Curriculum designers thus face a challenging question: How can mathematics instruction engage secondary students? Technology-based learning environments that adapt to student preferences may be an important vehicle to improve students understanding of and attitudes towards mathematics. However, designing these systems to "know" learners' lives, experiences, and preferences, and to utilize this knowledge in a helpful way, has been challenging. Accordingly, recent research in AI has focused on how to model learner characteristics using data from a variety of sources to provide engaging and adaptive learning experiences (Santos et al. 2016). Student preferences are one such type of data. Accounting for the different experiences, goals, and backgrounds of students through personalization and individualized pathways is cited as an important focus for those interested in artificial intelligence in education (Roll and Wylie 2016).

Here we test an intervention that utilizes *context personalization* (now "personalization" for short) – an instructional approach where curricular elements are matched to students' interests in areas like sports, movies, or games (Cordova and Lepper 1996). Positive effects have been found for AI-based systems that personalize reading passages (Heilman et al. 2010) as well as mathematics problems (Walkington 2013). We examine factors that moderate the success of a context personalization intervention that takes place within algebra instruction on linear relationships. Our research questions examine: (1) how performance and engagement differ in personalized versus standard curricular materials, (2) how a task-level factor – the *depth* of personalized connections made to students' interests – impacts performance and engagement, and (3) how a student-level factor – the degree to which students tend to quantitatively engage with their interest areas – moderates these effects. We find that both the degree to which students tend to engage quantitatively with their interest areas and the degree to which problem tasks are written to reflect how students actually engage with their interest areas are important factors in an AI environment enacting personalization.

Theoretical Framework

Individual and Situational Interest

Individual interests are the enduring predispositions that people have towards particular objects, events, and ideas (Hidi and Renninger 2006). Engaging with individual interests is associated with self-regulated learning behavior and metacognitive control, including self-reflection and the asking of curiosity questions (Hidi and Ainley 2008; Renninger and Su 2012; Sansone et al. 2011). A typical adolescent spends 7.5 h each day engaging with media – television, music, social networks, and video games (Rideout et al. 2010). Many adolescents also engage deeply with sports and the arts (Eccles et al. 2015). Walkington et al. (2014) identified common areas of interest for adolescents that have quantitative connections – including sports, video games, social networking, cell phones, other media, part-time jobs, and after-school activities.

When students have high levels of individual interest in a particular area, they often value that area and have positive feelings towards it (Renninger and Su 2012), and may engage with the area frequently, voluntarily, and independently (Renninger and Pozos-Brewer 2015). In this way, referencing or embedding tasks within the context of an individual interest through personalization may trigger students' *situational interest* – interest activated by salient and immediate characteristics of the environment (Hidi and Renninger 2006). Situational interest can spur affective reactions like increased enjoyment and value, as well as persistence and focused attention. A variety of studies (Høgheim and Reber 2015; Reber et al. 2009; Bernacki and Walkington 2018) have suggested that personalization can trigger situational interest.

In addition to having affective and value-laden reactions to their individual interest areas, importantly, students also tend to have significant background knowledge of an area for which they have individual interest (Renninger and Su 2012). This background knowledge can potentially be utilized to learn an academic subject by allowing for connections between new knowledge and prior knowledge. This phenomenon has been referred to as "grounding" in the cognitive science literature (e.g., Goldstone and Son 2005), and as "funds of knowledge" in the mathematics education literature (e.g., Civil 2007). Renninger and Hidi (2016) describe that, in order for learners' interest in a subject like mathematics to develop, they need to make a sustained connection to the content and be supported to seriously engage. Leveraging of knowledge related to out-of-school individual interests may be an important way to accomplish this.

In an AI environment, this triggered situational interest may be associated with students' engagement with the computer-based system – which can be measured by variables like students' tendency to "game the system" (Baker et al. 2004). Students have been observed to game problems in AI environments by quickly and systematically testing potential answers until the correct one is found, or by abusing the on-demand hint system by quickly cycling through hints to obtain the answer and then entering it. The persistence and focused attention that is spurred by situational interest may decrease the chance that learners engage in gaming behaviors. Gaming behaviors can be detected through data mining techniques (Romero et al. 2010), which are common for data from intelligent tutoring systems. Detectors can measure students' cognitive and affective engagement using log data. Baker and de Carvalho's (2008) gaming the system detector estimates each student's gaming tendency by examining errors made on the current problem as compared to previous problems, the speed of actions that follow commission of errors, whether the problem format allows for rapid, repeated answer selection, and whether knowing the skill is likely based on prior performance (see Baker et al. 2004). Baker and de Carvalho (2008) conducted a series of studies to examine how affect and motivation relate to gaming in ITS settings and found positive correlations between gaming the system and both a dislike of the subject matter (i.e., math) and a "lack of self-drive" (p. 219).

Researchers also have developed tools to detect attentional states related to learning in AI environments, including boredom and engaged concentration. Situational interest is described as moments of focused attention to stimuli which may or may not be sustained over time (Hidi and Renninger 2006), thus situational interest is related to and has the potential to modify students' attentional states. Baker et al.'s (2010) *affective states detectors* predict attentional states of boredom and concentration using indicators like previous help requests, duration of actions, how often a hint is followed by an error, actions where the student is predicted to have learned after guessing, giving an incorrect answer rather than asking for a hint, and so on. These detectors were developed using classroom observations of students' affective states (Ocumpaugh et al. 2012) and subsequent data mining methods to identify common features that appear in students' log files. Similar detectors developed by Baker and colleagues have successfully predicted learning outcomes (Aleven et al. 2006), including preparation for future learning (Baker et al. 2011), standardized test scores (Pardos et al. 2013) and college attendance (San Pedro et al. 2013). Another element related to attention is efficiency – Renninger and Hidi (2016) describe how processing interesting information is more efficient due to automatic allocation of attention. Koedinger et al. (2012) argue that because instructional time is so important in classrooms, the efficiency of learning is an important measure. In the present investigations, cognitive and affective processes that have hypothesized relations to situational interest and that are common to measure in an AI environment – such as efficiency, gaming, and affective states – are included.

Context Personalization: Incorporating Individual Interests

Based on the observation that many students hold strong individual interests in popular culture areas like sports, but less interest in mathematics, researchers have explored how school mathematics can become connected to students' individual interest areas outside of math. The approach of making connections to individual interests in math problems has been referred to as context personalization (Cordova and Lepper 1996). The hope is that by activating the positive affect, knowledge, and value that are associated with students' individual interests in topics like movies or music in the mathematics classroom, we can leverage these individual interests to support engagement with and conceptual understanding of mathematics. However, the results from research studies attempting to apply this principle have been disparate.

One common approach to personalization is to take a standard mathematics story problem and simply swap out words to match with students' background and interests. For example, studies have explored inserting students' names, favorite foods, pets, popular music artists, and other interesting-catching textual features into mathematics story problems. They compare students who receive such problems to students who receive standard problems, both in immediate performance on the modified problems and in post-test performance. Some very early studies suggested promise for this approach to personalization (Cordova and Lepper 1996; Davis-Dorsey et al. 1991), perhaps because of the novelty of inserting personalized identifiers into problems using efficient computer systems. More recent research, however, has found little effect for this "Mad Libs" style of personalization (Bates and Wiest 2004; Cakir and Simsek 2010; Høgheim and Reber 2015; Ku and Sullivan 2000; Simsek and Cakir 2009). Similarly, recent research on the Intelligent Tutoring System (ITS) *MATHia*, which sometimes selects problems based on students' ratings of interests in 4 broad areas (sports & fitness, the environment, business & money, music & art) actually found a slight negative correlation between interest-matched problems and measures of student performance, progress, and learning within the tutor, when examining a large sample of middle school students (Fancsali and Ritter 2014). And finally, Walkington et al. (2013) and Walkington et al. (2015b) found that making very simple and superficial interest-based modifications to 7th grade math problems, such as writing a problem about calculating the length of the diagonal of pop star Selena Gomez's closet, had little effect on students' performance, but did trigger interest.

It is not surprising that we do not see impressive results for this research. Inserting the name of one music artist versus another into a mathematics problem will do little to activate the rich, quantitative ways that students may engage with their interest areas. In addition, when considered through the lens of interest theory, this result is also not surprising – the connection to students' individual interests can be fleeting and relatively meaningless. Walkington and Bernacki (2014) proposed a framework for analyzing and interpreting the results of disparate personalization studies that focuses on how precisely the personalization is implemented. This framework captures the idea that personalization can be implemented at differing *depths* with respect to students' individual interests. Personalization can be designed to draw upon the ways in which students actually use number and quantities while pursuing their interest areas, such that an activation of students' prior knowledge of their interest area can potentially enhance their mathematical reasoning. This could be accomplished through carefully developed tasks situated in experiences that adolescents interested in a particular topic may have (Turner et al. 2012), like buying items at the store. Deep personalization is intended to draw upon students' prior knowledge in meaningful ways and incorporate how they actually pursue and participate in their individual interests.

Conversely, the personalization could be implemented at a surface level, where the quantitative reasoning involved has little to do with how students use numbers in their lives, and involves mathematics that students would never actually do. For example, Gutstein (2006) describes how the low-income Latino/a middle school students he taught reacted to the mathematics story problems in their curriculum. The students observed that, for instance, they don't go on canoe trips or ride in hot air balloons, they don't measure everything they see, and they don't spontaneously go downtown and count cars.

Depth differs from another aspect of Walkington and Bernacki's (2014) framework – personalization's *grain size*. Grain size captures the degree to which the particular interests of an individual student are accounted for. Personalization can be enacted at a fine grain size, where, for example, students interested in particular hockey teams get problems about those hockey teams. Or it can be implemented at a broad grain size, where these students are lumped into a "sports" category and get a variety of problems about playing and watching various sports. Depth focuses on the degree to which problems are connected to how math is "typically" used in the interest area by the target population, while grain size captures whether specific, individual students engage with the particular individual interest being described in the problem. The intervention here has a broad grain size while varying depth. Conversely, the "word-insert" method of personalization described earlier often has a fine grain size but very low depth. We next discuss research on personalization that explores ways to create deeper personalization.

Walkington (2013) created an intervention within the Cognitive Tutor Algebra (CTA) ITS where for one unit, 145 Algebra I students either received the standard algebra story problems already in the unit, or mathematically matched problems that were selected to correspond to their interests. The interest-based problems had been written based on surveys and interviews with students at the school site that attempted to measure both how they typically engaged with interest areas like sports or video games, and how they used numbers and quantities while engaging. All problems were focused on linear functions, which naturally had many strong connections to the ways students track changes in quantities over time while pursuing their interests (see Walkington et al. 2014). Study results showed that students had higher overall performance on the personalized problems, in terms of successfully writing an algebraic equation to go with a story problem on their first attempt. Also, students receiving personalization were able to write these algebraic expressions in less time and had fewer instances of "gaming the system" than their counterparts in the control group.

In subsequent work (Bernacki and Walkington 2018), we employed latent and path modeling techniques to examine how personalization impacts the relationship between affective, motivational variables like situational and individual interest, and cognitive variables, like efficiency, performance, and long-term learning in the same ITS for Algebra I. Results from an analysis of 151 Algebra I students using the tutor over a school year suggest that temporary states related to engagement (i.e., triggered situational interest) directly impact performance, and also feed into students' development of individual interest in mathematics as a subject area. Critically, this individual interest supports students' long-term learning and understanding of mathematical ideas. Here we follow these studies' approach to measuring performance and engagement – examining triggered situational interest as a measure of engagement, as well as problem-solving accuracy and time spent in a tutoring system (i.e., efficiency) as a measure of performance.

The studies reported in Walkington (2013) and Bernacki and Walkington (2018) used problems that had variable depth with respect to how students actually quantitatively engage with their interests. Some problems were relatively accurate depictions of how students had described actually using quantities and change while pursuing their interests, while others were superficially personalized by simply inserting a term related to an interest area to give the impression of match to one's interest, while others were in between. Overall, the depth of connections made to interests were moderate, and the results were in turn more promising than the studies reviewed earlier that made only superficial or surface connections. However, in the analysis all the problems were lumped together, as they were not designed and written to be easily separated into discrete categories. These studies also did not take into account the idea that students' degree of quantitative engagement with their interests may moderate the effectiveness of a personalization intervention. In other words, students who spend a lot of time thinking quantitatively and doing mental calculations while playing video games or watching sports may benefit more from receiving personalized mathematics tasks that evoke similar reasoning than those who engage with these same interests without significant quantitative engagement. This may be a critical factor that has been unaddressed by previous research.

In another recent line of studies (Walkington and Bernacki 2015; Walkington and Hayata 2017), middle school students engage in activities where they pose, share, and solve algebra problems relating to their interests. Qualitative analyses of students engaging in problem-posing suggest three key factors are important when considering the effectiveness of students' use of their individual interests when problem-posing: (1) students having an appropriate level of prior mathematical knowledge, (2) students grasping the norms and practices surrounding how mathematics story problems "work" in the classroom (e.g., must be solvable), and (3) the degree to which students engage quantitatively with their interests and actually use numbers to think about their interest areas. The third point is especially important. Some students who like sports, for example, may not engage with this interest area quantitatively – they may not closely track player or team statistics or the mathematical relationships that govern gameplay. They may not possess the prior knowledge related to sports that would be helpful to draw upon in the mathematics classroom when presented with sporting contexts. Thus, the degree of quantitative engagement with individual interest areas may be an important moderator of the effectiveness of personalization, but this factor has not been systematically investigated.

Two Key Factors for Personalization

Based on the above review, there are two important factors to consider when integrating students' individual interests into technology systems that have not been addressed in prior work. These considerations include characteristics of the learner and of the personalized tasks, each of which may moderate outcomes. We describe each of these in turn.

When students' individual interests are integrated into a computer-based curriculum, it may be important to acknowledge that students differ in the extent to which their prior knowledge of their interest area contains quantitative schema. For example, two students might each spend a lot of time playing video games, but they don't necessarily engage mathematically with the mechanics of the game in the same way. One student in Walkington et al. (2012) described how when playing video games "There's stuff like, this unit has 1000 health and does 100 damage per attack. And then the other units have they might have 10,000 health and they might to 20 damage per attack. If I have them attack each other, who will win?" Clearly, this student had rich quantitative engagement with the algebraic ideas that undergirded the game mechanics. Another student in the same study had a different level of engagement with the video game he played: "There's not really too many numbers, but sometimes there are numbers, but I haven't really… I can't really think of any..." Although the first-person shooter game being discussed in the latter quote does contain mathematical reasoning, this student did not report deep quantitative engagement while playing the game. These differences in students' knowledge about the mechanics of their individual interest area may have important implications. Personalization may be more effective for students with deeper quantitative prior knowledge – thus level of engagement might moderate the effect of personalization on performance.

In addition, when designers of online curricula attempt to bring students' individual interests into the classroom, they make decisions about what interests to draw upon and at what depth. Take the following problem written by a middle school student in Walkington and Bernacki (2015) about his favorite video game: "In Grand Theft Auto 5, you have 6 million dollars, and you want to buy the 2-type which cost 10 million dollars. How many missions will you have to do in order to buy the 2-type? Also you make 2.25 million each 1 missions." This problem is situated in a real quantitative reasoning context that students experience as they play video games – how long they will need to play to achieve certain rewards and advancements. It is plausible that students may be able to usefully draw upon their knowledge related to video gameplay to help them solve the problem. Contrast this with the following video game problem used in an intelligent tutoring system intervention reported in Walkington (2013): "You work at a store that sells computers and electronics. Your store just received 50 new Wii remotes, and you sell two each day. How many Wii remotes remain unsold after 16 days?" It is unlikely that students interested in video games spend time pondering how many Wii remotes a store sells each day. Although students may attend to the topic of this problem that involves video games, it is doubtful that they would be able to usefully apply their gaming-specific knowledge to the mechanics of this problem, which are focused on sales and not games.

Research Purpose and Questions

In the present study, we explore how the degree to which students have quantitative knowledge of their individual interests moderates the effectiveness of a context personalization intervention situated within an intelligent tutoring system. We examine an intervention that specifically varied the quantitative depth at which interests are incorporated into problem tasks – some problems were written to make surface level connections to interest areas, whereas others made more substantial quantitative connections. We hypothesize that problems that are written to meaningfully draw upon students' quantitative experiences in their everyday lives are more effective for student learning than those that make only surface connections. However, this effect may be moderated by whether the student has deep quantitative engagement with their interest area – these deeply-written problems may only be effective for students who have a high degree of quantitative engagement with their interests. Our research questions are as follows:

1. A. How does performance and engagement differ when students solve problems *personalized* to their interests compared to *standard* story Problems?

B. How is this effect moderated by the *degree of students' quantitative engagement* with their interest areas?

2. A. How does the *depth of the personalized connections* made to students' interests in problem tasks impact performance and engagement?

B. How is this effect moderated by the *degree of students' quantitative engagement* with their interest areas?

Note that research question 1a is a replication of prior research, while research questions 1b, 2a, and 2b are novel. Replicating previous work, particularly in an area where results have varied, is important. However, in our discussion and implications section, we focus mainly on the results of Research Question 2, as the depth manipulation is the novel contribution of this study.

Methods

Environment and Intervention

The personalization intervention took place within the ITS CTA software, an adaptive computer-based ITS for algebra (Carnegie Learning 2016). CTA uses AI to implement model-tracing approaches to relate problem-solving actions to a cognitive model of domain principles to give individualized feedback. CTA also uses knowledge-tracing approaches to track students' performance on problems in order to identify appropriate tasks to advance the student's knowledge (Koedinger and Aleven 2007). Much of CTA focuses on story problems (i.e., problems involving "real world" situations). Here we focus on Unit 9 which is titled "Linear Models and Ratios." This unit has story problems that model linear functions of the form $y = \mathrm{m}x$, where m is a simplified fraction with a denominator that is not equal to 1. There were three experimental conditions: (1) a control condition where students received the standard story problems from the standard version of CTA, (2) a "surface" personalization condition where students received 1 of 3 surface personalized versions of each problem selected by the software based on an interest survey, and (3) a "deep" personalization condition where students received 1 of 3 deeply personalized versions of each problem selected by the software based on an interest survey. The difference between surface and deep problems, as well as the interest survey instrument, are described fully in subsequent sections. Research question 1a was addressed by comparing condition (1) to conditions (2) and (3). Research question 2a was addressed by comparing condition (2) to condition (3). Research questions 1b and 2b were addressed by including degree of quantitative engagement with individual interests as a moderator.

Students were randomly assigned to one of the three conditions by CTA at the beginning of the school year; however, Conditions 2 and 3 only saw different experiences in the tutor when they reached Unit 9. Unit 9 was the fourth unit in a larger personalization intervention being implemented. In the previous 3 units impacted by the overarching personalization intervention, there was no surface versus deep distinction (i.e., no difference between Conditions 2 and 3), and all students in either condition had received personalized problems from the same problem bank. The problems in this bank had been a mix of problems that were surface, deep, or in between. As shown in Fig. 1, comparisons between Conditions 2 (surface personalization) and 3 (deep personalization) are not impacted by this overarching intervention, as they had identical prior experiences. However, comparisons between Condition 1 (control) and Conditions 2 and 3 combined (any personalized) may be impacted by the overarching intervention, in that any differences detected here may in part be due to the accrual of an effect based on work in these prior units. However, our main purpose here is to look at how the degree of students' engagement with individual interests and depth of personalization of the problems moderate outcomes, rather than formulate exact estimates for how different dosages of personalization impact performance. In other words, although the control versus personalized comparisons for research question 1a might be amplified or reduced as a result of students' prior history, a precise estimate of an effect size for personalization when compared to a control group is not the primary purpose of this study – instead, we focus on the moderation effect of degree of quantitative engagement and the surface versus deep problem comparison.

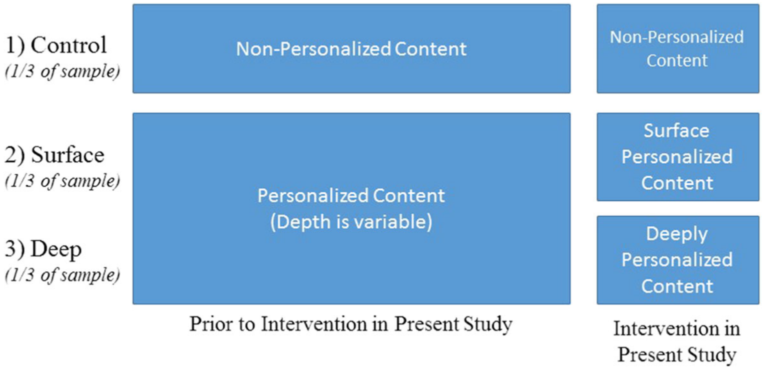


Fig. 1 Depiction of prior personalization intervention

Participants

The participants included $N = 106$ Algebra I students (58 female, 44 male, 4 unknown) using CTA at a suburban/rural high school in the Mid-Atlantic United States. The school served a student population that was 96% Caucasian with 21% of students eligible for free/reduced price lunch. In mathematics, 75% of students at the school passed the standardized state End-of-Course Exams. There were a total of 155 Algebra I students at the school, however this study only examines the 106 students who made it to Unit 9 in CTA before the school year ended. The percentage of students who made it to Unit 9 was nearly identical for the control group (70.6%) versus the personalized groups (70.7%), suggesting that the prior exposure to the overarching personalization intervention shown in Fig. 1 did not impact curriculum progress. Algebra I is a course typically taken in the eighth or ninth year of schooling in the United States (ages 13–15); however, students in the study were taking it in their ninth year (ninth grade) or later. Of the 106 students, 55 were taking Algebra 1 over one school year, 19 were taking Algebra I over 2 schools years and were in their first year (referred to as Algebra 1a), and 32 were taking Algebra I over two school years and were in their second year (referred to as Algebra 1b). Students were in 7 different classrooms of two Algebra I teachers, one of whom taught 6 classes (2 Algebra 1a, 2 Algebra 1b, 2 Algebra I) and the other taught one class (Algebra 1). Students were randomly assigned to the three conditions, with 36 in the Control, 35 in Surface Personalization and 35 in Deep Personalization.

Measures

Writing Surface and Deep Personalized Problems

We conducted two rounds of surveys ($N = 60$ and 45) and interviews ($N = 29$ and 23) with Algebra I students at the school site in years prior to the study concerning how they use numbers while pursuing their interests, and problems were written based on this data. The three interest categories used (sports, video games, and food) were selected based on this data as the most generative for algebraic connections. We found in these interviews, for example, that adolescents do keep track of sports statistics as well as their progress and probabilities relating to video games. Some examples are shown in Table 1.

Table 1 Transcripts of students discussing how they engage with interest areas

Sports	**Social networking**
Interviewer: Okay…So when you are doing track and cross country, um, do you use any numbers or see any numbers when you are doing them? *Student: Um, they have like distance marks, for like first mile, second mile, third, and then they'll have like arrows… and things like, there's tracking, um, how long it took you, what time it is, and everything like that, your speed…* *Interviewer: So do you know what your speed is?* *Student: Um, I get…usually get 25 min for 3 miles…*	*Interviewer: Do you see or use any numbers when you're using on Twitter or Instagram?* *Student: Not really…just like all my followers, like it shows you how many tweets you have…like your followers…stuff like that* *Interviewer: So, how many tweets do you have… around?* *Student: A lot…like 4000* *Interviewer: And do you keep track of how many you get all the time…or?* *Student: Yeah, it will…it like shows you when you go on, and it'll tell you and it'll tell you if you have new followers or you lost any followers…*
Video games	**Cooking**
Interviewer: What games do you play? *Student: I play Call of Duty, and Minecraft…* *Interviewer: How are numbers used in Call of Duty?* *Student: Like the kills, they'll have uh …the ratios like kills to death, stuff like that…* *Interviewer: So how many people you've killed* versus *how many times you've died?* *Student: Yeah, so like they use ratios in that…*	*Student: I want to be a chef, so the measurement things, like how to convert them...I'll definitely have to know those.* *Interviewer: So converting like what?* *Student: Like cups to quarts, times and stuff…*

Problems were written such that all three conditions had problems with identical algebraic and mathematical structures and numbers, and the problems were also matched on the punctuation, sentence structure, use of pronouns, use of mathematical vocabulary, etc. (see Walkington et al. 2015a, for a discussion of readability factors in algebra story problems). However, they varied in the depth of the connections made to students' interests (Table 2). In the control condition, students received problems that were written by the creators of CTA that were generally intended to be "personally or culturally relevant to students" (Koedinger 2001, p. 11). In the surface personalization

Table 2 An example of problem cover stories from each condition

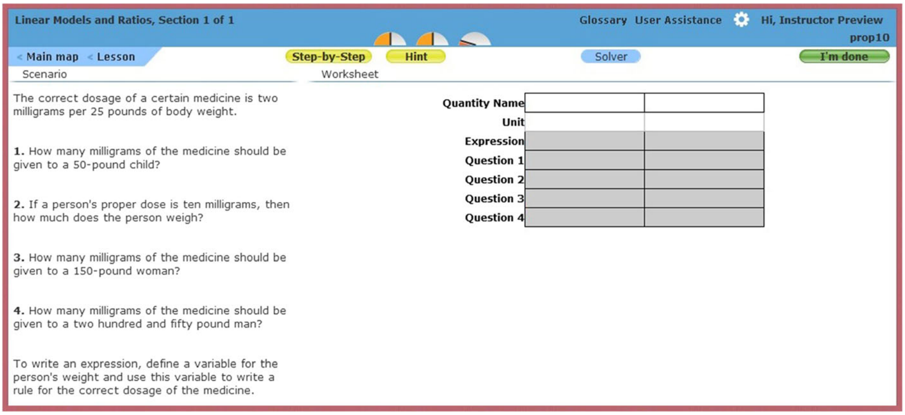
Condition	Example problem 1 (video games)	Example problem 2 (sports)
Control	You are a product inspector for a company that produces light bulbs. You find that two out of every 300 bulbs are defective: they don't work properly.	Three out of every five people in a recent survey supported the President's Health Plan.
Surface personalization	You are a product inspector assigned to the Gamestop chain of stores. You find that two out of every 300 locations have a violation: they are selling defective products.	Three out of five people have attended a Pittsburgh Steelers game in their lifetime.
Deep personalization	You enjoy playing World of Warcraft on your computer. You notice that two out of every 300 times you defeat a monster, the monster has an epic item: a treasure that you want to collect.	Three out of five free throws are successful for NBA players.

condition, students received problems about either sports, video games, or food that were written to be about their interest area, but that were not intended to invoke the ways adolescents might actually use numbers when engaging with that interest. For example, in Table 2, it is unlikely any adolescents think about GameStop's inspection policies or the attendance rates for Steelers' games as they pursue their interests. In the deep personalization condition, students received problems about sports, video games, or food that were written to correspond to ways they actually might use numbers and algebra while pursuing their interests in these areas.

There were 11 problems in the original unit; however, due to the tutor's model-tracing approach, not all students received all problems, and students could receive problems twice (usually with different numbers) if they were struggling. Alongside the "cover stories" given in Table 2, the students were asked to name the quantities and units in the problem, write an algebraic expression for the problem, and solve for particular x and y cases (see Fig. 2). For the analyses here, we focus on students writing the algebraic expression rather than other steps where students were solving for particular x and y cases. We focus on expression-writing because it is the underlying concept in the problem that is truly "algebraic" (rather than using arithmetic), and it is where we would expect students to be directly engaging with the story context itself (rather than simply plugging numbers into an expression they have already written). Expression-writing, along with connecting algebraic expressions to verbal and tabular representations, is a critical skill in the Common Core State Standards (CCSS 2010) for Algebra and for Functions. Specifically, the CCSS include standards about seeing the structure in expressions, writing equations and building functions to model relationships between quantities, interpreting functions that model mathematics in context, and analyzing functions using different representations (e.g., tabular, symbolic). CTA collects detailed log-files of all student interactions with the system; these files were uploaded into the DataShop platform (Koedinger et al. 2010).

Interest Surveys

To assess students' interests in different topics, we administered interest surveys where students numerically rated their interest in, time spent engaging, and how much they



The correct dosage of a certain medicine is two milligrams per 25 pounds of body weight.

1. How many milligrams of the medicine should be given to a 50-pound child?

2. If a person's proper dose is ten milligrams, then how much does the person weigh?

3. How many milligrams of the medicine should be given to a 150-pound woman?

4. How many milligrams of the medicine should be given to a two hundred and fifty pound man?

To write an expression, define a variable for the person's weight and use this variable to write a rule for the correct dosage of the medicine.

Quantity Name		
Unit		
Expression		
Question 1		
Question 2		
Question 3		
Question 4		

Fig. 2 Screenshot of problem in Unit 9 of Cognitive Tutor Algebra software

knew about different topics, as well as responded to an open-ended prompt. Before entering Unit 1 in CTA, participants were given this interests survey where they rated their interest in 10 topics – music, art, cell phones, food, computers, games, stores, TV, movies, and sports. They responded to the prompt "Rate your level of interest in (topic)" for each of the 10 areas on a 1–4 scale (It's Boring, It's Okay, I Like It, It's my Favorite Thing). The responses to these items were used to assign students to problem versions in the personalization conditions. For students in the personalization condition, the average rating on food was 3.34 ($SD = 0.76$), the average on sports was 2.97 ($SD = 1.05$), and the average on games was 2.88 ($SD = 0.91$). Within the personalization conditions, if the student's highest rating was equal in two interest areas, the problem version was randomly selected from among these 2 versions.

After each prompt to rate their engagement with an interest topic, students would also respond to the open-ended prompt, "Describe how you use numbers when you are doing things related to this interest" for each of the 10 topics. This question was used to measure degree of quantitative engagement with interest areas. Student responses to this prompt were coded as follows. A *0* indicated that they did not answer, gave an uninformative response, or gave an example of using numbers that involved simply seeing numbers (e.g., on phone buttons) or engaging in counting or measuring (e.g., counting points scored in a game). A *1* indicated that they discussed adding quantities or comparing quantities in the context of their interest. A *2* indicated that they discussed multiplicative (i.e., algebraic) relationships in the context of their interests, including ratios, proportions, scale factors, rates, percents, and so on. A random sample of $N = 110$ student responses was coded on the 0–2 scale by a second coder, who obtained a weighted kappa reliability of 0.80. See Appendix A for more information on the coding.

For each student, we used the maximum rating they received for any of the 10 interest areas on the 0–2 scale as our measure of quantitative engagement with interests. This was intended to be an overall measure of how deeply they tend to engaged quantitatively with their interest areas. We use the term "engagement" rather than "knowledge" because we asked students to describe the things they did while pursuing their interests, rather than formally assessing their quantitative knowledge. The maximum value was used because of the relative rarity of a students' response getting a 2 rating.

Students were also asked how much time they spent engaging with each interest topic and how much they knew about each interest topic; these variables were not related to the effect of Condition on our outcome variables, so are not considered here further. As part of the larger study, we also collected pre-measures of students' situational and individual interest in mathematics using the Linnenbrink-Garcia et al. (2010) scales. These scales allowed us to assess how interested students were in the domain of mathematics at the beginning of the school year, using items like "I enjoy doing math" and "What we are studying in math class is useful for me to know." These variables were also not related to the effect of Condition on our outcome variables, so are also not considered here further.

Performance Measures

We used two measures of performance within Unit 9. First, the proportion of correct first attempts on expression-writing was calculated for each student in Unit 9. The correctness on *first attempts* is typically used as a performance measure in CTA because students eventually get every problem correct after making additional attempts informed by correctness feedback and any hints they request. Second, the number of algebraic expressions written correctly per minute was also calculated for each student as a measure of learning efficiency. The metric corrects per minute does count instances where the student achieved the correct answer while using the hint system in CTA as correct. We wanted to include these instances because we generally consider appropriate use of the help system to be a desirable behavior – students should seek feedback when they need it, and effectively use that feedback to help them get and understand the correct answer. However, the help system can certainly be abused (e.g., by clicking through the hints and just copying the answer in the "bottom out" hint) – this is why we also include a measure of gaming the system, described next.

Detectors

We used Baker and de Carvalho's (2008) gaming detector (described earlier) to estimate gaming tendencies (scaled from 0 to 1), which were averaged for each student across expression-writing steps in the unit. We also used Baker et al.'s (2010) affective state detector (described earlier). The predictions of bored, confused, frustrated, and concentrating behaviors were averaged for each student across expression-writing steps in the unit, and were also on a 0–1 scale. We narrowed our consideration to only concentrating and bored affective states, as frustrated and confused behavior estimates had near zero variance in Unit 9.

The gaming detector was trained using data from students utilizing Cognitive Tutor Algebra (Baker and de Carvalho 2008), using an approach shown to generalize across students and Cognitive Tutor lessons (Baker et al. 2008b).[1] Although gaming and affect values ranged from 0 to 100, the values used in the analysis were not rescaled to be interpretable as estimates of the percentage of transactions in the tutor where students engaged in these behaviors. Thus, when interpreting the size of the effects, we rely on Cohen's *d* calculations, as the raw coefficients do not afford ready interpretation in meaningful terms.

Situational Interest Pop-Ups

Finally, we measured students' triggered situational interest in the problems. For every other problem, students in all conditions would receive a pop-up prompt which read "This problem was interesting" and that requested a rating on a scale of 1 (Not Really)

[1] Both of these studies took place with students in the same geographic area and of a similar age group as the present study. The gaming detector had also been validated using 8th and 9th grade students from the same geographic area using a Cognitive Tutor for 8th grade math designed by the same company that designed Cognitive Tutor Algebra (Carnegie Learning Inc.; Baker et al. 2005; Baker et al. 2004). The affective states detector had been validated using students in the same grade level, in the same area, using the same ITS system (Cognitive Tutor Algebra), around the same time period, in Baker et al. (2012). The unit the detector was validated for involved story problems with systems of linear equations instead of story problems with single linear equations, but was similar. The affective states detector was also validated with high school students using a unit on scatter plots in Cognitive Tutor Algebra (San Pedro et al. 2014), and in studies with other tutors that cover algebra content and that involved 8th grade or high school students (Baker et al. 2010; Pardos et al. 2013).

to 4 (Very). Each student's responses to these prompts were averaged over the unit; 9 students (4 control, 2 surface personalized, 3 deep personalized) were omitted from these analyses because they chose to never respond to this prompt.

Analysis

In the present study, we examined (1a) the effect of personalization to out-of-school interests on student performance and engagement, (1b) how this effect was moderated by degree of quantitative engagement with out-of-school interests, (2a) the effect of the depth of personalization on these outcomes, and (2b) how this effect was moderated by degree of quantitative engagement with out-of-school interests. We fit linear regression models to the data using the *lm ()* command in the R software program. The outcome measures were: (1) proportion of correct first attempts, (2) correct equations written per minute, (3) tendency to "game the system" on a scale of 0–1, (4) ratings of problem interestingness on a 1–4 scale, and (5) tendency to display bored or concentrating behaviors on a scale from 0 to 1. Multi-level models that nested students in classrooms and/or teachers were initially fit; however, the random effect variance for these nesting variables was near 0, and results were the same either way. Thus we moved forward with the simpler models. Experimental condition was included as a predictor in each model.

Control variables were included for whether the student was in Algebra 1, 1a, or 1b, as well as their course grade in mathematics the prior year. No prior course grade was available for 11 students; thus it was imputed using the Amelia package in R. Amelia uses multiple imputation with bootstrapping and the Expectation-Maximization algorithm to impute missing values (Honaker, King, and Blackwell 2011). The imputed math grades for the 11 students had a mean that was close to the mean for the remaining 95 students (*t*-test for differences non-significant at $p = 0.87$), and the standard deviations were also similar, suggesting that these students did not differ from the rest of the sample. Amelia assumes a multivariate normal distribution of the variable to be imputed. Although course grades did not follow a normal distribution in their raw form, when exponentiated to the 5th power they were not significantly different than normal (via Shapiro-Wilks). Results of all analyses were nearly identical whether course grade was in its original or exponentiated form prior to normalization, so we keep the variable on its original scale for interpretability. Course grade was normalized by subtracting the mean for the 95 (106–11) students from each value, and dividing by the standard deviation from across the 95 students. Additional analyses were run with prior standardized tests and/or a paper-based unit test as controls; however, results were not sensitive to the decision of which prior achievement measure to use, and course grades had the smallest number of missing values.

For our first research question (impact of personalized versus control problems) we collapsed the two personalization conditions (surface and deep). For our second research question, we only examined the subset of the data of students who received personalized problems, and contrasted surface-level personalization with deep-level personalization. Because students in both of these personalization conditions had received problems from the same problem bank in the prior unit on writing expressions from story problems, we were able to additionally include a control for performance in the prior unit.

Students' degree of engagement with their interests was also a predictor in the model, as coded on a 0–2 scale previously described. This variable was specified as an ordered factor variable in R using the *ordered()* command. All main effects were included in each model. Then all possible 2-way interactions were systematically tested for inclusion using the *anova()* command on nested models; this command uses a chi-square distribution to test for significant reductions in model deviance.

In the present analysis, given our relatively small sample size and the exploratory nature of the study, instead of relying solely on *p*-values to guide our discussion, we use multiple measures of impact and significance. In particular, we combine the presentation of 95 % confidence intervals for all regression coefficients, with ΔR^2 measures of effect size from comparisons of regression models, with Cohen's *d* effect sizes from the raw data, with *p*-values, to obtain a holistic sense of the results.

Results

Descriptive Statistics

The first five rows in Table 3 show descriptive statistics on the control/background variables. There were no significant differences between groups using chi-squared tests (for math course) and *t*-tests (for math grade and degree of engagement). The remaining rows are outcome variables. Overall, students achieved the correct answer on their first attempt to write a linear eq. 70.09% of the time in Unit 9 ($SD = 18.04\%$), and on average produced 3.18 ($SD = 2.01$) correct answers per minute. The students' average rating of the interestingness of the problems on the 4-point scale was 2.17 ($SD = 0.96$). On a 100-point scale, average gaming was 22.77 in the unit ($SD = 14.21$), and average affective state measures were 85.77 ($SD = 17.98$) for boredom and 43.30 ($SD = 8.21$) for concentration.

Table 3 Descriptive statistics for control, background, and outcome measures

Measure	Control ($N = 36$)	Surface ($N = 35$)	Deep (N = 35)
No. students taking Algebra 1 (%)	18 (50.00%)	19 (54.29%)	18 (51.43%)
No. students taking Algebra 1a (%)	6 (16.67%)	6 (17.14%)	7 (20.00%)
No. students taking Algebra 1b (%)	12 (33.33%)	10 (28.57%)	10 (28.57%)
Normalized prior math grade (SD)	0.02 (1.00)	0.34 (0.94)	0.17 (1.07)
Avg. degree quantitative engagement (SD)	0.47 (0.51)	0.60 (0.50)	0.57 (0.50)
No. students in low engagement category (%)	19 (52.78%)	14 (40.00%)	15 (42.86%)
No. students in medium engagement category (%)	9 (25.00%)	13 (37.14%)	12 (34.29%)
No. students in high engagement category (%)	8 (22.22%)	8 (22.86%)	8 (22.86%)
% Correct first attempts unit 9 (SD)	65.33% (16.33)	74.80% (11.49)	70.29% (23.48)
Correct per min unit 9 (SD)	2.52 (1.10)	3.95 (2.09)	3.09 (2.41)
Avg. problem interestingness rating unit 9 (SD)	1.86 (0.92)	2.11 (0.96)	2.55 (0.92)
Avg. gaming tendency unit 9 (SD)	28.60 (13.58)	18.72 (13.47)	20.81 (13.98)
Avg. concentration unit 9 (SD)	45.33 (7.31)	40.47 (8.29)	44.05 (8.46)
Avg. bored unit 9 (SD)	82.53 (16.42)	92.93 (11.09)	81.94 (22.77)

Research Question 1: Performance and Engagement on Personalized Vs. Normal Problems

To examine the overall effect of personalization, we first conducted analyses that collapsed the two personalization conditions (surface and deep) and compared this collapsed condition to the control condition. The control and combined personalization groups had similar average scores on our normalized prior grade measure (Average for Control = 0.02, *SD* = 1.00; Average for Personalized = 0.026, *SD* = 1.00; *t*-test difference not significant at $p = 0.25$). Model selection warranted the inclusion of a Condition by Degree of Interest Engagement interaction term in the model predicting corrects per minute ($\chi^2(1) = 6.16$, $p = 0.015$). All other two-way interactions were non-significant.

Correct First Attempts

As can be seen from Table 4, the personalization condition did not significantly outperform the control condition on proportion of correct first attempts. However, this result was close to significance ($p = 0.076$), and we would likely have replicated the results in Walkington (2013) with a slightly larger sample. Students who received personalization answered algebraic expression writing problems with 6.69% (95% CI [−0.70%, 14.07%]) greater accuracy. This corresponded to a raw effect size of Cohen's $d = 0.41$, which is consistent with prior results for the impact of personalization on performance which suggests an effect size around $d = 0.28$ (see Walkington 2013).

Table 4 Linear regression analyses comparing personalized to control condition ($N = 106$)

Predictors	Model 1: prop. correct B (SE) $^{\text{Sig}}$	Model 2: correct/min B (SE) $^{\text{Sig}}$	Model 3: gaming B (SE) $^{\text{Sig}}$	Model 4: problem interestingness B (SE) $^{\text{Sig}}$
Intercept	0.679 (0.035) ***	3.016 (0.365)***	0.277 (0.027)***	1.921 (0.193)***
Algebra 1	(ref.)	(ref.)	(ref.)	(ref.)
Algebra 1a	−0.035 (0.048)	−0.620 (0.486)	0.012 (0.037)	0.194 (0.260)
Algebra 1b	−0.058 (0.041)	−1.354 (0.415)**	0.031 (0.032)	−0.158 (0.231)
Prior math grade	0.012 (0.018)	0.112 (0.186)	−0.013 (0.014)	−0.064 (0.100)
Control condition	(ref.)	(ref.)	(ref.)	(ref.)
Personalization condition	0.067 (0.037)	1.158 (0.503)**	−0.085 (0.029)**	0.451 (0.204)*
Degree of interest engagement	0.004 (0.031)	−0.263 (0.528)	0.017 (0.024)	0.315 (0.172)
Personalization x interest Engagement		1.635 (0.659)*		
ΔR^2 for condition main effect	3.04%	4.01%	7.99%	4.85%
ΔR^2 for condition interaction		4.72%		

* = $p < .05$, ** = $p < .01$, *** = $p < .001$. (ref.) denotes the reference category for factor variables. The reference category for numerical variables is simply "0." However, the reference category for degree of interest engagement is the middle category, or "1." Note that prior math grade is normalized. Coefficient values for proportion correct need to be multiplied by 100 to become percentages

Corrects per Minute

A main-effects-only model suggested that the personalized group had 0.86 more corrects per minute than the control (95% CI [0.10, 1.62], $p = .028$). A model that allowed for interactions (shown in Table 4) suggested that for students who were at medium (level 1) for degree of engagement with their interests (the reference category in the table), the personalization condition had 1.16 more corrects per minute than the control condition (95% CI [0.38, 1.94], $p = .004$). The interaction term reveals that the personalization group outperformed the control group even more when students had a high degree of engagement with their interests (level 2), corresponding to an additional 1.64 correct expressions written per minute (95% CI [0.33, 2.94]). However, students with a low degree of engagement with their interests (level 0) did not outperform the control group, achieving directionally fewer (1.16–1.64 = −0.48) corrects per minute. The overall difference between personalized and control in corrects per minute corresponded to a raw Cohen's d of 0.56. For students with engagement with their interests at level 1 or 2 specifically, this effect size rose to 0.92. Students in the Algebra 1b track also had significantly fewer corrects per minute than students in the Algebra I class (1.35 fewer; 95% CI [0.53, 2.18], $d = 0.76$, $p = .002$).

Engagement Measures: Gaming, Affect, and Interest

The gaming detector analysis estimated the personalization group was less likely to engage in gaming behaviors than the control group ($p = .002$), which corresponds to a raw Cohen's d of 0.65. This replicates the result in Walkington (2013) which found a significant gaming reduction for the personalization condition with a d effect size of 0.35. Finally, students in the personalization condition rated problems as significantly more interesting – approximately 0.45 points higher (95% CI [0.05,0.86]) on a 4-point scale ($p = 0.029$). This corresponds to a raw Cohen's d of 0.50. There was also a marginally significant positive effect ($p = 0.071$) of degree of interest engagement on problem interestingness rating, such that each level of engagement corresponded to a 0.32 increase (95% CI [−0.03, 0.66]) in interest rating. This corresponded to a raw effect size of $d = 0.54$ for the difference between a level 0 and level 2 students' rating.

Additional analyses were conducted using the affective states detector; however, there were no significant differences between the personalization and control condition (ps > 0.1). Overall, there is evidence that the personalization condition outperformed the control condition on some measures of performance and engagement, particularly when degree of quantitative engagement with interests was high.

Research Question 2: Performance and Engagement of Surface Versus Deep Personalization

For our next analysis, we restricted the sample to only those students who had received personalization ($N = 70$ students), and compared the 35 students who had received surface personalization to the 35 students who had received deep personalization. On our normalized prior grade measure, the average for the deep personalization condition was 0.35 ($SD = 1.07$) and the average for the surface personalization condition was 0.17 ($SD = 0.94$; t-test difference nonsignificant at $p = 0.46$). Model selection warranted the inclusion of a Personalization Condition by Degree of Interest Engagement interaction in the model for proportion of correct first attempts ($\chi^2(1) = 4.00$, $p = .0499$) and a Degree of Interest Engagement by Course interaction for Problem Interestingness Ratings ($\chi^2(2) = 3.65$, $p = .032$). No other 2-way interactions were significant.

Correct First Attempts

In a main-effects-only model, no significant differences were found between the surface and deep personalized groups for correct first attempts ($p = 0.50$). In a model that allowed for interactions (see Table 5), there was no effect of deep versus surface personalization on correct first attempts when students' degree of quantitative engagement with their interests is medium (at level 1, which is the reference category in the table; $p = 0.96$). However, for students who engage with their interests to a greater degree (level 2), the deep personalization condition significantly outperforms the surface personalization condition by 16.5% on correct first attempts (95% CI [0.00%, 33.04%], $p = .0499$), corresponding to a raw effect size of $d = 0.39$. Alternately, for students who did not engage with their interests in a quantitative manner (level 0), the performance of the surface personalization group is actually significantly higher than the deep personalization group by a similar amount (0.003–0.165 = 0.162, or 16.2%), corresponding to a raw effect size of $d = -0.43$. Further, the table shows that proportion of correct firsts in the prior unit predicts proportion of correct firsts in Unit 9 – which amounted to a 34.73% (95% CI [4.04%,65.42%], $p = 0.027$) predicted difference for a student who had gotten no first attempts correct in the previous unit, compared to a student who had gotten all first attempts correct.

Table 5 Linear regression analyses comparing deep personalized condition to surface personalized condition ($N = 70$ students)

Predictors	Model 1: prop. correct B (SE) ^{Sig}	Model 2: correct/min B (SE) ^{Sig}
Intercept	0.580 (0.092)***	3.14 (0.779)***
Algebra 1	(ref.)	(ref.)
Algebra 1a	−0.015 (0.059)	−0.075 (0.647)*
Algebra 1b	−0.095 (0.052)	−1.693 (0.572)**
Prior math grade	0.003 (0.023)	0.155 (0.245)
Prior unit performance on outcome measure	0.347 (0.154)*	0.872 (0.379)*
Degree of interest engagement	−0.082 (0.059)	1.007 (0.455)*
Surface personalization condition	(ref.)	(ref.)
Deep personalization condition	−0.003 (0.046)	−0.909 (0.468)
Deep personalization x interest engagement	0.165 (0.082)*	
ΔR^2 for condition main effect	0.65%	3.98%
ΔR^2 for condition interaction	5.49%	

* = $p < .05$, ** = $p < .01$, *** = $p < .001$. (ref.) denotes the reference category for factor variables. The reference category for numerical variables is simply "0." However, the reference category for degree of interest engagement is the middle category, or "1." Note that prior math grade is normalized. Coefficient values for proportion correct need to be multiplied by 100 to become percentages

Corrects per Minute

There was a marginally significant p-value of $p = 0.057$ for a negative effect of deep personalization on correct answers per minute, compared to surface personalization (0.91 fewer corrects per minute, 95% CI [−0.03,1.85]), corresponding to a d of -0.38). This suggests that deep personalization may involve less learning efficiency. For each additional level of quantitative engagement, correct answers per minute increased for either type of personalized problem (1.01 more correct answers per minute per level, 95% CI [0.10,1.92], $p = .030$). The raw effect size for the difference between high and low interest engagement on corrects per minute was $d = 0.75$. However, as we saw in the analysis for research question 1, degree of engagement with interests did not affect corrects per minute for the Control condition. Thus students who have high quantitative engagement with their interests may benefit from receiving any kind of personalized problem. Having achieved more correct answers per minute in the prior unit is associated with more correct answers per minute in Unit 9 ($p = 0.025$). Finally, students in Algebra 1b had 1.69 (95% CI [0.55, 2.84], $d = 0.93$, $p = 0.004$) fewer correct answers per minute than students in Algebra 1.

Engagement Measures: Gaming, Affect, and Interest

There were no variables that significantly predicted gaming. However, as can be seen from Table 6, students who received deep personalization were less likely to be bored ($p = .021$) than students who received surface personalization, with a raw d effect size of 0.61. Students receiving deep personalization were directionally more likely to be

Table 6 Linear regression analyses comparing deep personalized condition to surface personalized condition ($N = 70$ students)

Predictors	Model 3: gaming B (SE) ^{Sig}	Model 4: concentrating B (SE) ^{Sig}	Model 5: boredom B (SE) ^{Sig}	Model 6: problem interestingness B (SE) ^{Sig}
Intercept	0.096 (0.068)	0.210 (0.110)	0.824 (0.078)***	0.453 (0.214)
Algebra 1	(ref.)	(ref.)	(ref.)	(ref.)
Algebra 1a	−0.026 (0.047)	0.013 (0.026)	−0.024 (0.060)	0.452 (0.225)*
Algebra 1b	0.042 (0.040)	0.058 (0.024)*	−0.066 (0.052)	−0.312 (0.181)
Prior math grade	−0.019 (0.018)	−0.012 (0.010)	0.018 (0.022)	0.033 (0.075)
Prior unit performance on measure	0.283 (0.197)	0.361 (0.224)	0.258 (0.136)	0.853 (0.089)***
Degree of interest engagement	0.029 (0.030)	−0.00 (0.18)	−0.049 (0.041)	0.0181 (0.182)
Surface personalization condition	(ref.)	(ref.)	(ref.)	(ref.)
Deep personalization condition	0.027 (0.033)	0.038 (0.019)	−0.101 (0.043)*	−0.156 (0.158)
Algebra 1a x interest engagement				0.445 (0.416)
Algebra 1b x interest engagement				−0.620 (0.294)*
ΔR^2 for condition main effect	1.00%	5.05%	7.29%	0.55%

* = $p < .05$, ** = $p < .01$, *** = $p < .001$. (ref.) denotes the reference category for factor variables. The reference category for numerical variables is simply "0." However, the reference category for degree of interest engagement is the middle category, or "1." Note that prior math grade is normalized

concentrating than those receiving surface personalization. This difference did not pass the threshold for significance ($p = 0.053$), but corresponded to a raw d effect size of 0.43. In addition, Algebra 1b students were slightly more likely to be concentrating than Algebra 1 students ($d = 0.66$, $p = 0.015$). Although a slightly larger sample would yield more definitive results, there is evidence that deep personalization has a positive impact on students' affective states compared to surface personalization. Further, neither degree of engagement with interests nor the interaction of degree of engagement and condition was significant for the affective states models ($p\text{s} > 0.2$), indicating a consistent effect of depth of personalization.

For problem interestingness ratings, students who rated problems as more interesting in the prior unit were more likely to rate them as more interesting in this unit (0.85 points higher for every average point higher in previous unit, 95% CI [0.67, 1.03]). Finally, the interaction term for problem interestingness ratings suggests that for students in Algebra 1b, having a higher level of quantitative engagement with their interests is associated with lower problem ratings, compared to students in Algebra I (0.62 points lower per level, 95% CI [0.03, 1.21]) or Algebra 1a (1.07 points lower per level, 95% CI [0.18, 1.95]).

Overall results suggest that deep personalization is more effective for students who have considerable quantitative engagement with their interests, while surface personalization is more effective for students who report engaging with the quantitative aspects of their interests at a superficial level. Further, deep personalization seems to have some positive benefits for all students in terms of suppressing negative affective states.

Sensitivity Analysis

As issues with statistical power are present throughout our results, we present a sensitivity analysis to find the minimum detectable effect size for our given sample. The size of the sample was pre-determined in our context in that every Algebra I student at the school participated in the study. Using a β of .80 to allow us to have 80% chance of detecting the effect, and an α of .05 to allow a 5% chance of obtaining a false positive, we calculated the size of the d-type effect detectable with our current sample restrictions. The G*Power 3.0.10 software was used (Faul et al. 2009). For the comparison of control versus personalized, an effect size of d = 0.58 would be needed to detect effects in our sample. For the comparison of surface versus deep personalization, an effect size of d = 0.68 would be needed to detect differences in our sample. However, as can be seen throughout our results section and as summarized in Table 7, our effect sizes were smaller than these minimums (typically around d = 0.40–0.50). This is consistent with prior research on personalization – effects are generally in the small to medium range (e.g., Walkington 2013; Bernacki and Walkington 2018). Thus, personalization research in general could benefit from larger sample sizes, which may not always be possible in studies that involve single schools.

Summary of Results

Results are summarized in Table 7. Note that given our holistic approach, we include findings with marginally significant p-values; however, we italicized the findings that did not reach traditional measures of statistical significance.

Table 7 Summary of Findings Relating to Condition

Outcome variable	Finding	*d* effect size
Correct first attempts	• *Students who receive personalization have more correct first attempts than the control group.*	0.41
	• Students who receive surface personalization have more correct first attempts than deep personalization when their degree of engagement with their interests is low (level 0).	-0.43
	• Students who receive deep personalization have more correct first attempts than surface personalization when their degree of engagement with the interests is high (level 2).	0.39
Corrects per minute	• Students who receive personalization have more correct answers per minute than the control group if they have higher engagement (level 1 or level 2) with their interests)	0.92
	• *Students who receive deep personalization have fewer corrects per minute than those who receive surface personalization.*	-0.38
Gaming the system	• Students who receive personalization are less likely to game the system than the control group.	0.65
Problem interestingness rating	• Students who receive personalization rate problems as more interesting than the control group.	0.50
Concentrating behavior	• *Students who receive deep personalization are more likely to be concentrating than those who receive surface personalization.*	0.43
Bored behavior	• Students who receive deep personalization are less likely to be bored than those who receive surface personalization.	0.61

Findings that did not reach a 0.05 p-value for significance are indicated in italics

Discussion

In the present study, we examined two important moderating factors for the effectiveness of personalization interventions – degree of students' knowledge about their interest area, and depth of the interest-based connection made in the written problem. Our major findings were that (1) personalized problems showed an overall advantage over standard problems, (2) the effects of surface versus deep personalization on performance were moderated by students' degree of engagement with their interests, (3) deeper personalization tended to elicit lower efficiency but more positive affective states, while situational interest and gaming were not impacted, and (4) students who engaged more deeply with their interests were stronger on efficiency measures. We discuss each of our results in turn.

Personalized Versus Standard Story Problems

Our first line of inquiry examined differences between students receiving personalized problems versus standard story problems across measures of performance and engagement. In prior work (Walkington 2013), we found that students who receive personalization have higher performance and efficiency on personalized problems. Here, we replicated this result and observed that it appeared to be somewhat moderated by students' degree of quantitative engagement with their interests. The present study also replicated previous results that students receiving personalization are less likely to game the system (Walkington 2013) and report higher triggered interest (Bernacki and Walkington 2018) than their counterparts in the control group. However, as these results are somewhat limited by the effect of the prior intervention on detected differences, we turn to discussing our primary contrast of interest – surface versus deep personalization.

Surface Versus Deep Personalization – Performance and Efficiency Measures

Our second inquiry explored the differences between performance and engagement on surface versus deep personalized problems. In terms of correct first attempts, we found that for students who report greater quantitative engagement with their interests, deep personalization seemed more effective, while for students who report less quantitative engagement with their interests, surface personalization seemed more effective. It is important to note that findings from our first research question suggest that *both* types of personalization are, however, as good as or better than standard story problems. The result that students who have more quantitative engagement with their interests may benefit more from problems designed to tap into deep quantitative interest features was expected. These problems were created to draw upon a particular knowledge base that only these students have, thus it is not surprising these students would do better on them. However, the finding that the surface personalized problems were significantly better for students who reported engaging with their interests at a superficial level was more surprising. It might be hypothesized that for these students, whether the problem is surface or deep does not matter, and that any interest-based connection in their mathematics problems will trigger the affective components related to their interest area (but not the knowledge components). This result suggests that if the problem is written to have deeply personalized features, it might actually be disruptive to these students when compared to a problem that is still personalized, but does not have any of the deep quantitative features. This is supported by the more general finding that students in the deep personalization condition had fewer corrects per minute.

Durik and Harackiewicz (2007), found that adding colorful pages with visually stimulating fonts and pictures increased motivation for students who had low individual interest in math (IIM), but hampered motivation for those with high IIM. Conversely, a deeper intervention that accentuated communicating the value of learning math to everyday activities had the opposite effect for both groups – it supported motivation for those with high IIM, but dampened it for low IIM. Although here the effect of personalization is moderated by quantitative engagement with interests rather than IIM, the finding is similar – for students with less quantitative engagement, a shallower intervention improved in-tutor performance and a deeper intervention was detrimental, while for those with more quantitative engagement, the opposite was true.

An example to illustrate this is the deeply personalized problem “A website shows that in the game Minecraft just one out of every five grass blocks that a player hits will give grass seeds.” Students who reported a low level of quantitative engagement with their interests sometimes mixed up the independent and dependent variables in the equation, whereas those who reported a higher level of quantitative engagement with their interests did not. To someone not familiar with Minecraft, it would make sense that grass seeds cause grass blocks. However, someone who plays Minecraft would know that grass blocks actually cause grass seeds – you harvest the seeds. Thus, the deeply personalized features of this problem might require particular domain-specific knowledge of video games to understand.

Surface Versus Deep Personalization – Engagement Measures

The pattern was slightly different for eliciting affective states. In particular, we found that deep personalization led to more positive affective states for all students, regardless of their prior knowledge of their interest area. This suggests that the richness and realism of these problems is more attractive to students and elicits more positive affect, even though the prior knowledge or cognitive requirements for solving these problems may be greater. Students may be unaware that they lack the subtle kinds of background knowledge that may be needed for deeply personalized problems, and may enjoy solving the more relevant problems even in the face of limited success in doing so. This raises an important question: does decreasing boredom help students to come to understand mathematics better? This is an important area to investigate - for example, in one recent study, Walkington et al. (2016) found that adding illustrations to problems actually depressed post-test performance compared to a condition that received text only, even though students rated that they liked the problems with illustrations more. This relates to research on *seductive details* (Harp and Mayer 1998; Schraw and Lehman 2001) – extraneous information added to learning materials to enhance interestingness, that has been shown disrupt students' learning by increasing the cognitive demand. Accordingly, the *coherence principle* (Mayer 2009) suggests that removing interesting but irrelevant information should foster learning. Whether the interest area-specific information contained in deep personalization is always a seductive detail is unclear, however, given that this information can be quantitative in nature and has the potential to support students with deep knowledge of their interest area in understanding mathematical relationships.

We also found that although there was some evidence deep personalization elicited more positive affective states, there were not differences in situational interest between deep and surface personalized groups, nor did we find differences in gaming the system. We return to this finding in the implications section.

Overall Effect of Students' Degree of Interest Engagement

Finally, we found that students who report engaging more deeply with their interests tend to have better learning efficiency when solving any type of personalized problem, compared to students who engage with their interests at a surface level. It is not surprising that students who tend to think quantitatively in their everyday lives tend to be stronger problem-solvers. An interesting direction for future research might be to explore how we can get students to think more deeply about the quantitative rules that govern their everyday activities and experiences when they are outside of school. This links to the important Common Core Mathematical Practice of modelling problems arising in everyday life, society, and the workplace using mathematics (CCSS 2010).

Limitations

There are several limitations to the current study that must be acknowledged. First, the sample size was limited, resulting in insufficient power to meet thresholds for statistical significance with effects of a size similar to those obtained in past studies – this limits the strength of all the conclusions we can draw. The limited sample size became reduced further as students were separated into different groups based on learner characteristics. Second, our prior mathematics achievement measure (prior course grades) was quite distal to the outcomes examined here (performance in an individual tutoring unit), and thus did not have much predictive power. For our second research question, we were able to use the far more proximal measure of performance in a prior tutoring unit. And finally, the operationalization of students' degree of quantitative engagement with their individual interests, and surface versus deep personalized problems, were limited by the instructional context the study took place in – an intelligent tutoring system. For example, coding of open-ended interview data may have allowed us to better operationalize each students' degree of interest engagement. Similarly, giving students complex, open-ended problem-based scenarios might be a better operationalization of "deep" personalization. Although the ITS context is limited in some respects, it is an efficient, scalable, and tightly-controlled way to conduct investigations into student engagement and learning. It also offers fine-grained log data of student interactions with the system, which were leveraged here.

In addition, we find it important to acknowledge the interactive nature of research questions, methods, and results in the present study. In particular, because of the larger study this intervention was a part of, we had a variety of student background variables available, each of which we discussed in our Methods section. We had measures of prior achievement (grades, standardized test, paper test) and prior interest in mathematics (situational, individual). Research would suggest that math interest and math achievement may moderate the effect of interventions like personalization (Walkington 2013; Durik and Harackiewicz 2007). We also described quantitative measures we had regarding how much students "liked" different interest topics, how much time they spent on them, and how much they reported knowing about them, in addition to our degree of quantitative engagement measure. Prior research suggests that the extent of students' engagement with their out-of-school interests might be a moderator of personalization as well (Walkington and Bernacki 2015).

When conducting our initial analyses, once we had established that Condition (personalization vs. control) seemed to matter in its main effect, we sought to explore whether the effectiveness of personalization differed based on our background characteristics. We tested all of our background variables both for main effects and for interactions with Condition at some point during our preliminary analyses, and were fascinated to find that the only one that mattered consistently was degree of quantitative engagement with interests. This, in turn, drove the framing of the current study, and the narrowed research focus and method. Thus, this approach is exploratory, and has some obvious limitations, including a considerable inflation of the risk of a Type I error. However, the idea that quantitative engagement with interests is a powerful construct in predicting math learning is an important one, particularly since its contribution appears to be distinct from math achievement, math interest, and general engagement with interests. The notion that this idea is valuable is what drove the present study and analysis; however, ultimately replication must be sought.

Implications

We present implications in two areas – for the theory of personalization and for the design of AI curricular environments. First, a theoretical contribution of this study is the direct contrast of surface versus deep approaches to personalization. Although previous studies have used one approach or the other (e.g., Cordova and Lepper 1996; Walkington and Bernacki 2014), we were able to systematically compare the approaches against each other to understand their influence on motivation, cognition, and learning. Results suggest that the story is not as simple as the tempting adage that "deeper personalization is better" – instead, it may depend on the characteristics of the learner. Indeed, Renninger and Hidi (2016) describe that although interest relates to both the characteristics of an individual and the characteristics of an environment, "much of the research focuses on one or the other of these two aspects… although they perhaps should be considered together" (p. 8). This study shows the potentially powerful types of findings that can come from this simultaneous consideration. However, further studies that directly contrast the two approaches and take into account learner characteristics are needed to corroborate findings.

A related theoretical point is the finding that the degree of quantitative engagement with out-of-school interests mattered for student performance, both in its main effect and as a moderator of personalization. This variable mattered when more traditional measures identified in the literature – like individual interest in mathematics and prior mathematics grades (Durik and Harackiewicz 2007; Walkington 2013) – did not. This suggests that students' quantitative engagement with their "funds of knowledge" (Moll, Amanti, Neff, and Gonzalez 1992) – the home and community-based knowledge resources they gain outside of school that are numerical or measured – may be an important factor in student success in mathematics. This also suggests that the "knowledge" component of individual interest (Renninger and Su 2012), which is not always accounted for in typical interest measures (e.g., Linnenbrink-Garcia et al. 2010), may be critical to understanding the implications of engagement with out-of-school individual interests.

A final and related theoretical point is that personalization tends to spur situational interest compared to typical instruction (e.g., replication of Høgheim and Reber 2015), but no further distinctions in situational interest were observed between deep and surface personalized conditions. Situational interest is a multidimensional construct that includes the triggering and maintenance of interest by activating affective as well as cognitive processes (Renninger and Hidi 2016). In other research focused on processes within interest development, we distinguish between the ways that experiences of enjoyment and perceptions of value contribute to the effects of personalization (Bernacki and Walkington 2018). In this study, detector analyses allow us to further examine processes. Results indicate that, within a group of students who already experienced greater situational interest due to personalization of problems, students further demonstrated greater concentration and less boredom when problems made deeper versus more superficial connections to their prior knowledge about their interest. The presence of differences in situational interest between control and personalized groups and the lack of difference within deep and surface conditions suggests that additional distinctions in the sources of situational interest may need to be drawn. Detector analyses represent a first step in disentangling such cognitive-affective processes and highlight that, even amongst those with elevated interest in a task, additional dimensions of personalization can further induce desirable differences in engagement in problem-solving (i.e., heightened concentration and decreased boredom) but that the deeper inclusion of prior knowledge might come at a cost (i.e., to learning efficiency, where accessing prior knowledge necessary to solve deep problems may increase cognitive demand). It is also interesting that these affective differences did not translate into gaming differences between the surface and deep conditions – instead, the gaming findings tracked closely with the situational interest findings. Processes that spur gaming the system may be closely linked but oppositional to those that trigger situational interest, congruent with research suggesting gaming's association with disliking math (Baker et al. 2008a, b).

Conclusion

Personalizing learning to students' interests using adaptive technology systems has the potential to improve student performance and efficiency, elicit interest, and induce more positive affective states. In mathematics specifically, presenting students with problems in the context of interests like sports and video games is shown to be beneficial. As Renninger and Hidi (2016) argue, supporting learners to make connections between the learner and the content (here, mathematics) is critical to interest development, which can lead to improved learning. However, the impact of such personalization interventions is not uniform, and based on the data reported here seems to be moderated by two important factors. First, students' engagement with their interest area – how much they tend to engage quantitatively with sports or video games in their everyday lives – seems to be a key factor. In addition, the degree to which the problem is written to realistically invoke relevant quantitative schemes relating to an individual interest also moderates outcomes. The identification of these two key moderators for personalization interventions has important implications for curriculum designers seeking to implement personalization effectively in their computer-based systems.

Acknowledgements This work was supported by the Pittsburgh Science of Learning Center, which was supported by the National Science Foundation, Award #0836012. Thank you to Gail Kusbit, Ryan Baker, Steve Ritter, Steve Fancsali, Vincent Aleven, and Timothy Nokes-Malach for their generous assistance with this research. Thank you to Carnegie Learning for their collaboration and support of this project. Thank you to Elizabeth Howell and Alyssa Holland for their assistance with data transcription and coding. And finally, thank you to the classroom teachers involved in this study.

Appendix A

Coding of Degree of Quantitative Engagement with Interests

Figure 3 shows the percentage of student responses coded as Level 0, Level 1, and Level 2. The majority of student responses were coded at Level 0, indicating that students did not make quantitative connections to their interests, made surface-level connections to seeing numbers, or made connections simply to counting or measuring.

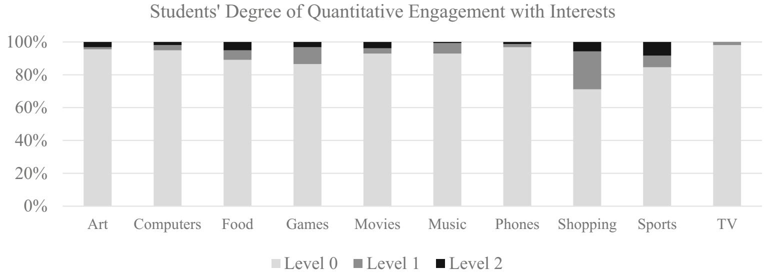


Fig. 3 The percentage of student responses to the prompt "Describe how you use numbers when you are doing things related to this interest" that were coded as Levels 0–2

Some examples included responses like "I do not use numbers while using the computer," "dialing phone numbers," and "I use numbers to count how many songs I am downloading on my phone." Typically, between 2% and 10% of responses were at Level 1, where the student described comparing or adding quantities. Some examples included "to add up all the costs of the things you buy to know how much money you're going to spend," "you use numbers when you want to know how many calories and figuring how much you want to eat," and "to keep track of who is winning the game." Finally, between 0% and 7% of responses were at Level 2, where the student is making some sort of algebraic multiplicative comparison or conversion. Some examples included "you have to calculate everything, the cost of a large scale army to the probability of hitting the opponent and whether or not you can take the risk," "drawing to scale," and "if you are baking or cooking and you to find measurements and convert them." As can be seen from Fig. 3, the most generative topics for Level 1 and 2 connections were Sports, Food, Games, Shopping, and Movies. Further, we found that 48 students in the sample reported engaging with all categories at Level 0 (45%), 34 students (32%) reported engaging with at least one topic at Level 1 but no topics at Level 2, and 24 students reported in engaging with at least one topic at Level 2 (22%).

References

Aleven, V., Mclaren, B., Roll, I., & Koedinger, K. (2006). Toward meta-cognitive tutoring: A model of help seeking with a cognitive tutor. *International Journal of Artificial Intelligence in Education, 16*(2), 101–128.

Baker, R., & de Carvalho, A. (2008). Labeling student behavior faster and more precisely with text replays. In R. S. J. de Baker, T. Barnes, & J. E. Beck (Eds.), *Educational data mining: 1st international conference on educational data mining, proceedings* (pp. 38–47). Montreal, Quebec, Canada: International Educational Data Mining Society.

Baker, R.S., Corbett, A.T., Koedinger, K.R., Wagner, A.Z. (2004) Off-task behavior in the cognitive tutor classroom: When students "game the system." *Proceedings of ACM CHI 2004: Computer-Human Interactions*, 383–390.

Baker, R. S., Corbett, A., Koedinger, K., & Roll, I. (2005). Detecting when students game the system, across tutor subjects and classroom cohorts. *Proceedings of User Modeling, 2005*, 220–224.

Baker, R.S.J.d., Corbett, A.T., Roll, I., Koedinger, K.R. (2008a) Developing a generalizable detector of when students game the system. User Modeling and User-Adapted Interaction, 18, 3, 287–314.

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

Baker, R. S. J.d., D'Mello, S. K., Rodrigo, M. M. T., & Graesser, A. C. (2010). Better to be frustrated than bored: The incidence, persistence, and impact of Learners' cognitive-affective states during interactions with three different computer-based learning environments. *International Journal of Human-Computer Studies, 68*(4), 223–241.

Baker, R.S.J.d., Gowda, S., Corbett, A.T. (2011) Towards predicting future transfer of learning. *Proceedings of 15th International Conference on Artificial Intelligence in Education*, 23–30.

Baker, R.S.J.d., Gowda, S.M., Wixon, M., Kalka, J., Wagner, A.Z., Salvi, A., Aleven, V., Kusbit, G., Ocumpaugh, J., Rossi, L. (2012) Towards sensor-free affect detection in cognitive tutor algebra. *Proceedings of the 5th International Conference on Educational Data Mining*, 126–133.

Bates, E., & Wiest, L. (2004). The impact of personalization of mathematical word problems on student performance. *The Mathematics Educator, 14*(2), 17–26.

Bernacki, M. L., & Walkington, C. (2018). The role of situational interest in personalized learning. *Journal of Education & Psychology, 110*(6), 864–881. https://doi.org/10.1037/edu0000250.

Cakir, O., & Simsek, N. (2010). A comparative analysis of computer and paper-based personalization on student achievement. *Computers & Education, 55*, 1524–1531.

Carnegie Learning. (2016). Cognitive Tutor Algebra. [software]. Pittsburgh, PA.

Civil, M. (2007). Building on community knowledge: An avenue to equity in mathematics education. In N. Nassir. and P. Cobb (Eds.) *Improving access to mathematics: Diversity and equity in the classroom* (pp. 105–117).

Common Core State Standards Initiative. (2010). Common Core State Standards (Mathematics Standards). Retrieved 6 July 2010 from http://www.corestandards.org/the-standards/mathematics

Cordova, D., & Lepper, M. (1996). Intrinsic motivation and the process of learning: Beneficial effects of contextualization, personalization, and choice. *Journal of Educational Psychology, 88*(4), 715–730.

Davis-Dorsey, J., Ross, S., & Morrison, G. (1991). The role of rewording and context personalization in the solving of mathematical word problems. *Journal of Educational Psychology, 83*(1), 61–68.

Durik, A. M., & Harackiewicz, J. M. (2007). Different strokes for different folks: How individual interest moderates the effects of situational factors on task interest. *Journal of Educational Psychology, 99*(3), 597–610.

Eccles, J., Fredricks, F., & Epstein, A. (2015). Understanding well-developed interests and activity commitment. In A. Renninger, M. Nieswandt, & S. Hidi (Eds,) *Interest in Mathematics and Science Learning* (pp. 315–330), American Educational Research Association, Washington D.C.

Fancsali, S. E., & Ritter, S. (2014, March). Context personalization, preferences, and performance in an intelligent tutoring system for middle school mathematics. In *Proceedings of the Fourth International Conference on Learning Analytics and Knowledge* (pp. 73–77). ACM.

Faul, F., Erdfelder, E., Buchner, A., & Lang, A.-G. (2009). Statistical power analyses using G*power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods, 41*, 1149–1160.

Fredricks, J. A., & Eccles, J. (2002). Children's competence and value beliefs from childhood through adolescence: Growth trajectories in two male-sex-typed domains. *Developmental Psychology, 38*, 519–533.

Frenzel, A. C., Goetz, T., Pekrun, R., & Watt, H. M. G. (2010). Development of mathematics interest in adolescence: Influences of gender, family, and school context. *Journal of Research on Adolescence, 20*, 507–537.

Goldstone, R. L., & Son, J. Y. (2005). The transfer of scientific principles using concrete and idealized simulations. *The Journal of the Learning Sciences, 14*(1), 69–110.

Gutstein, E. (2006). Reading and writing the world with mathematics: Toward a pedagogy for social justice. Taylor & Francis.

Harp, S. F., & Mayer, R. E. (1998). How seductive details do their damage: A theory of cognitive interest in science learning. *Journal of Educational Psychology, 90*(3), 414–434.

Heilman, M., Collins-Thompson, K., Callan, J., Eskenazi, M., Juffs, A., & Wilson, L. (2010). Personalization of reading passages improves vocabulary acquisition. *International Journal of Artificial Intelligence in Education, 20*(1), 73–98.

Hidi, S., & Ainley, M. (2008). Interest and self-regulation. The relationships between two variables that influence learning. In D. H. Schunk & B. J. Zimmerman (Eds.), *Motivation and self-regulated learning: Theory, research, and application* (pp. 77–109). New York: Erlbaum.

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

Honaker, J., King, G., & Blackwell, M. (2011). Amelia II: A program for missing data. *Journal of Statistical Software, 45*(7), 1–47. http://www.jstatsoft.org/v45/i07/.

Høgheim, S., & Reber, R. (2015). Supporting interest of middle school students in mathematics through context personalization and example choice. *Contemporary Educational Psychology, 42*, 17–25.

Kaput, J. J. (2000). *Teaching and learning a new algebra with understanding*. U.S.; Massachusetts: National Center for Improving Student Learning and Achievement.

Koedinger, K. (2001). Cognitive tutors as modeling tool and instructional model. In K. D. Forbus & P. J. Feltovich (Eds.), *Smart Machines in Education: The coming revolution in educational technology*. Menlo Park: AAAI/MIT Press.

Koedinger, K. R., & Aleven, V. (2007). Exploring the assistance dilemma in experiments with cognitive tutors. *Educational Psychology Review, 19*, 239–264.

Koedinger, K.R., Baker, R.S.J.d., Cunningham, K., Skogsholm, A., Leber, B., Stamper, J. (2010). A data repository for the EDM community: The PSLC DataShop. In Romero, C., Ventura, S., Pechenizkiy, M., Baker, R.S.J.d. (Eds.) *Handbook of Educational Data Mining*. Boca Raton: CRC Press.

Koedinger, K. R., Corbett, A. T., & Perfetti, C. (2012). The knowledge-learning-instruction (KLI) framework: Toward bridging the science practice chasm to enhance robust student learning. *Cognitive Science, 36*, 1–42.

Ku, H., & Sullivan, H. (2000). Personalization of mathematics word problems in Taiwan. *Educational Technology Research and Development, 48*(3), 49–59.

Linnenbrink-Garcia, L., Durik, A. M., Conley, A. M., Barron, K. E., Tauer, J. M., Karabenick, S. A., & Harackiewicz, J. M. (2010). Measuring situational interest in academic domains. *Educational and psychological measurement, 70*(4), 647–671.

Mayer, R. E. (2009). *Multimedia learning* (2nd ed.). New York: Cambridge University Press.

McCoy, L. P. (2005). Effect of demographic and personal variables on achievement in eighth-grade algebra. *Journal of Educational Research, 98*(3), 131–135.

Moll, L. C., Amanti, C., Neff, D., & Gonzalez, N. (1992). Funds of knowledge for teaching: Using a qualitative approach to connect homes and classrooms. *Theory into practice, 31*(2), 132–141.

Moses, R., & Cobb, C. (2001). *Radical equations: Math literacy and civil rights*. Boston: Beacon Press.

National Mathematics Advisory Panel. (2008). Foundations for success: The final report of the National Mathematics Advisory Panel. *US Department of Education. Retrieved from* http://www2.ed.gov/about/bdscomm/list/mathpanel/report/final-report.pdf.

Ocumpaugh, J., Baker, R. S. J. D., & Rodrigo, M. M. T. (2012). *Baker-Rodrigo Observation Method Protocol (BROMP) 1.0. Training Manual version 1.0. Technical report*. New York: EdLab. Manila, Philippines: Ateneo Laboratory for the Learning Sciences.

Pardos, Z. A., Baker, R. S. J.d., San Pedro, M. O. C. Z., Gowda, S. M., & Gowda, S. M. (2013). Affective states and state tests: Investigating how affect throughout the school year predicts end of year learning outcomes. *Proceedings of the 3rd International Conference on Learning Analytics and Knowledge*, 117–124.

Reber, R., Hetland, H., Chen, W., Norman, E., & Kobbeltvedt, T. (2009). Effects of example choice on interest, control, and learning. *The Journal of the Learning Sciences, 18*(4), 509–548.

Renninger, K., & Pozos-Brewer, R. (2015). Psychology of interest. In J. D. Wright (Ed.), *International Encyclopedia of the Social & Behavioral Sciences 2nd edition Volume 12* (pp. 378–385). Oxford: Elsevier.

Renninger, K., & Su, S. (2012). Interest and its development. In R. M. Ryan (Ed.), *Handbook of human motivation*. NY: Oxford University Press.

Renninger, K. A., & Hidi, S. (2016). The power of interest for motivation and engagement. New York, NY: Routledge.

Rideout, V. J., Foehr, U. G., & Roberts, D. F. (2010). Generation M [superscript 2]: Media in the Lives of 8-to 18-year-olds. Henry J. Kaiser Family Foundation. Retrieved from http://files.eric.ed.gov/fulltext/ED527859.pdf

Roll, I., & Wylie, R. (2016). Evolution and revolution in artificial intelligence in education. *International Journal of Artificial Intelligence in Education, 26*(2), 582–599.

Romero, C., Ventura, S., Pechenizky, M., Baker, R. (2010). Handbook of educational data mining. 2010. Editorial Chapman and Hall/CRC Press, Taylor & Francis Group. Data Mining and Knowledge Discovery Series.

San Pedro, M.O.Z., Baker, R.S.J.d., Bowers, A.J., Heffernan, N.T. (2013) Predicting college enrollment from student interaction with an intelligent tutoring system in middle school. *Proceedings of the 6th International Conference on Educational Data Mining*, 177–184.

San Pedro, M.O.Z., Baker, R.S.J.d., Mercedes, M.M.T. (2014) Carelessness and affect in an intelligent tutoring system for mathematics. *International Journal of Artifiical Intelligence in Education,* 24, 189–210.

Sansone, C., Fraughton, T., Zachary, J., Butner, J., & Heiner, C. (2011). Self-regulation of motivation when learning online: The important of who, why, and how. *Educational Technology Research and Development, 59*, 199–212.

Santos, O. C., Kravcik, M., & Boticario, J. G. (2016). Preface to special issue on user modelling to support personalization in enhanced educational settings. *International Journal of Artificial Intelligence in Education, 26*(3), 809–820.

Schraw, G., & Lehman, S. (2001). Situational interest: A review of the literature and directions for future research. *Educational Psychology Review, 13*(1), 23–52.

Simsek, N., & Cakir, O. (2009). Effect of personalization on students' achievement and gender factor in mathematics education. *International Journal of Social Science, 4*, 278–282.

Turner, E., Drake, C., McDuffie, A., Aguirre, J., Bartell, T., & Foote, M. (2012). Promoting equity in mathematics teacher preparation: A framework for advancing teacher learning of children's multiple mathematics knowledge bases. *Journal of Mathematics Teacher Education, 15*, 67–82.

Walkington, C. (2013). Using learning technologies to personalize instruction to student interests: The impact of relevant contexts on performance and learning outcomes. *Journal of Educational Psychology, 105*(4), 932–945.

Walkington, C., & Bernacki, M. (2014). Motivating students by "personalizing" learning around individual interests: A consideration of theory, design, and implementation issues. In S. Karabenick & T. Urdan (eds.) *Advances in Motivation and Achievement Volume 18* (pp. 139–176), Emerald Group Publishing.

Walkington, C., & Bernacki, M. (2015). Students authoring personalized "algebra stories": Problem-posing in the context of out-of-school interests. *The Journal of Mathematical Behavior, 40B*, 171–191.

Walkington, C., & Hayata, C. (2017). Designing learning personalized to students' interests: Balancing rich experiences with mathematical goals. *ZDM Mathematics Education, 49*(4), 519–530. https://doi.org/10.1007/s11858-017-0842-z.

Walkington, C., Sherman, M., & Petrosino, A. (2012). 'Playing the game' of story problems: Coordinating situation-based reasoning with algebraic representation. *The Journal of Mathematical Behavior, 31*(2), 174–195.

Walkington, C., Cooper, J., & Howell, E. (2013). The effects of visual representations and interest-based personalization on solving percent problems. In M. Martinez & A. Castro Superfine (Eds.), *Proceedings of the 35th annual meeting of the North American Chapter of the International Group for the Psychology of Mathematics Education* (pp. 533–536). Chicago: University of Illinois at Chicago.

Walkington, C., Sherman, M., & Howell, E. (2014). Personalized learning in algebra. *Mathematics Teacher, 108*(4), 272–279.

Walkington, C., Clinton, V., Ritter, S., & Nathan, M. J. (2015a). How readability and topic incidence relate to performance on mathematics story problems in computer-based curricula. *Journal of Educational Psychology, 107*(4), 1051–1074.

Walkington, C., Cooper, J., Nathan, M. J., & Alibali, M. A. (2015b). The effects of visual representations and interest-based personalization on solving mathematics story problems. In T. Bartell, K. Bieda, R. Putnam, K. Bradfield, & H. Dominguez (Eds.), *Proceedings of the 37th annual meeting of the North American Chapter of the International Group for the Psychology of Mathematics Education (p. 127)*. East Lansing: Michigan State University.

Walkington, C., Clinton, V., & Mingle, L. (2016). Considering cognitive factors in interest research: Context personalization and illustrations in math curricula. In M. B. Wood, E. E. Turner, M. Civil, & J. A. Eli (Eds.), *Proceedings of the 38th annual meeting of the North American Chapter of the International Group for the Psychology of Mathematics Education* (pp. 89–96). Tucson, AZ: The University of Arizona.