Computers & Education 237 (2025) 105347

Contents lists available at ScienceDirect

Computers & Education

journal homepage: www.elsevier.com/locate/compedu

Leveling the computational playing field: Inquiring about factors predicting computational thinking in constructionist game-based learning

Giovanni M. Troiano [a],*,[1], Amir Abdollahi [b],[1], Michael Cassidy [c], Gillian Puttick [c], Tiago Machado [d], Casper Harteveld [a]

[a] Northeastern University, Art + Design, 360 Huntington Ave, Boston, MA, 02115, United States
[b] Northeastern University, Mechanical and Industrial Engineering, 360 Huntington Ave, Boston, MA, 02115, United States
[c] TERC, 2067 Massachusetts Ave, Cambridge, MA, 02140, United States
[d] IBM Research, Av. República do Chile, 330. Centro Rio de Janeiro, RJ 20031-170, Brazil

ARTICLE INFO

Keywords:
Computational thinking
Digital literacy
STEM
Game-based learning
Constructionist learning
Scratch
Dr. Scratch
K-12
Multilevel modeling

ABSTRACT

Computational thinking (CT) is key in STEM and computer science (CS) education. Recently, there has been a surge in studies inquiring about the factors that predict the CT development of young students. We extend these prior works by inquiring about the factors that predict the CT of students (n = 932) in a constructionist game-based learning (GBL) STEM curriculum. Specifically, after addressing missing data through imputation, we apply Multilevel Modeling (MLM) to identify these potential factors in Scratch games and students' CT. We found that teachers' experience implementing game-based curricula, students' Scratch experience, student choice of game genre, and the interaction between teacher experience and game genre significantly predicted CT. Instead, students' gender did not emerge as a significant predictor of CT. We provide recommendations for curricula that support CT through constructionist GBL.

1. Introduction

Computational thinking (CT) is key to contemporary education (Barr & Stephenson, 2011). Nevertheless, implementing CT in school curricula remains challenging for various reasons, including a lack of consensus on CT definitions and standard assessment methods. Among other challenges, ongoing inquiries in CT education concern the factors that predict CT uptake. Amidst potential factors, there is prominent interest in how and if students' gender may predict CT (Román-González et al., 2017). Such inquiries support a discussion of gender representation in computer science (CS, Beyer 2014). Prior work in STEM showed how males' and females' CT is generally alike (Chongo et al., 2020; Hutchins et al., 2017). Other research found differences between genders, showing how female students showed a risk-averse, curiosity-driven, and disciplined attitude toward CT, while male students showed a risk-taking attitude and felt confident in their understanding of CT (Sovey et al., 2022). These inquiries considered other predictors, including computational creativity (Israel-Fishelson et al., 2021), programming self-efficacy (Kong et al., 2018), academic achievement (Sun et al., 2022), and teacher support (Jin et al., 2021). By revealing factors likely to predict CT development in young students, these inquiries provided critical insights that can be used to improve educational approaches to scaffolding CT

* Corresponding author.
E-mail addresses: g.troiano@northeastern.edu (G.M. Troiano), abdollahi.a@northeastern.edu (A. Abdollahi), michael_cassidy@terc.edu (M. Cassidy), gilly_puttick@terc.edu (G. Puttick), tiago.machado@nyu.edu (T. Machado), c.harteveld@northeastern.edu (C. Harteveld).
[1] These authors contributed equally to this work.

https://doi.org/10.1016/j.compedu.2025.105347
Received 25 April 2024; Received in revised form 25 April 2025; Accepted 4 May 2025

Available online 7 June 2025

strategically. Here, we inquire about factors predicting CT in constructionist game-based learning (GBL). We focus on constructionist GBL because designing games benefits CT the most among other project-based learning activities (Moreno-León et al., 2017). As part of a constructionist GBL STEM curriculum, 932 middle-school (i.e., 8th-grade) students designed games about climate change topics (e.g., ice-albedo feedback) in Scratch (Resnick et al., 2009). The factors under analysis, among others, include students' demographics, teachers' experience implementing the curriculum, and students' prior experience with Scratch.

Through Multilevel Modeling (MLM, Roberts 2004), we analyze a secondary dataset comprising 483 games designed by dyads (pairs) of 8th-grade students to gauge potential factors predicting CT, their hierarchical structure, and potential interactions between them. We employ MLM because it can effectively analyze data organized at multiple hierarchical levels (e.g., Khine 2022). This approach not only facilitates the examination of predictive factors, akin to Structural Equation Modeling (SEM, Markus 2012), but also clarifies the hierarchical relationships among these factors and determines the levels—whether group or individual—at which they exert influence. Our results are relevant to STEM and CS research, educators, and curriculum designers who wish to empower educators in teaching CT and further refine professional development (PD) for scaffolding successful CT learning.

1.1. CT in constructionist game-based learning

In a recent paper, Troiano et al. (2020b) described GBL as either (1) *instructionist*, characterized by a teacher-centered or direct instruction method, to impart knowledge and skills to students through structured lectures or (2) *constructionist*, which is learner-centered and encourages active engagement, hands-on activities, and emphasizes students as active creators (Honey, 2013). The curriculum providing the data for our MLM (Section 2.1) used GBL to have students construct knowledge of climate change and validate it through game design in Scratch (Resnick et al., 2009). In essence, GBL that leverages game design embodies the constructionist principles of "learn-by-creating" purveyed by Seymour Papert and underpinned by Jean Piaget's cognitive development theory (Papert, 1980; Piaget, 1959) and can greatly facilitate the development of problem-solving skills—thus of CT. Further, game design is shaped by *design thinking* (Lockwood, 2010), which maps well with the scope and practices of CT and can support CT uptake by fostering critical thinking, problem-solving, decision-making, and the nurturing of creativity (Kafai, 1995).

Recent work on CT via constructionist GBL (Troiano et al., 2020a) has observed that factors such as the game genre of student-designed games predict CT uptake and shape students' CT development. They found that storytelling games consistently scored lower than other game genres due to less frequent use of logical statements (e.g., *if*, *if-else*). However, the authors discussed how these results might inadvertently hinder female students, who generally prefer storytelling games to engage with game design and CT development in game-based learning (Troiano et al., 2020a)—a trend that has been discussed in the literature of CS education (Werner et al., 2009). These findings resonate with a surging wave of studies investigating factors predicting CT uptake, which call for further and similar inquiries to contribute to discussions relevant to CT, STEM, and broadening participation in CS (Peckham et al., 2007).

1.2. Factors predicting CT

Prior work has examined multiple factors predicting CT, where affective attitudes, personal attitudes, creative thinking, mathematical proficiency, socio-economic status, and resilience emerged as significant predictors (Atman Uslu, 2023; Durak & Saritepeci, 2018; Guggemos, 2021; Israel-Fishelson et al., 2021; Korucu et al., 2017; Moon & Cheon, 2023; Polat et al., 2021). Teaching methods and cultural contexts also predict CT, with studies showing the impact of physics curricula, peer collaboration, and age on developing CT skills (Kong et al., 2018; Latifah et al., 2022; Lei et al., 2020; Werner et al., 2012). Among CT's potential predicting factors, students' gender and its interaction with CT in STEM and CS education were prominent and analyzed through SEM, cross-lagged regression, and logistic regression (Atman Uslu, 2023; Durak & Saritepeci, 2018; Kong & Lai, 2022; Peugh, 2010; Relkin et al., 2020; Sun et al., 2022). MLM was not employed. Studies considering how and if gender predicts CT contributed to a broader discussion on representation in STEM and CS (Baser, 2013; Czerkawski & Lyman, 2015; Funke & Geldreich, 2017; Kiss, 2010; Sax et al., 2017; Seiter & Foreman, 2013; Serkan & Karalar, 2018; Stein, 2004; Stoilescu & Egodawatte, 2010; Werner et al., 2004). These studies produced mixed results, with some finding males more confident and proficient in computer-based CT tasks, while others found females excel more in collaborative and unplugged tasks, but most found that students' gender does not predict CT (Ardito et al., 2020; Del Olmo-Muñoz et al., 2020; Delal & Oner, 2020; Gao et al., 2022; Hutchins et al., 2017; Niousha et al., 2022; Rojas López & García-Peñalvo, 2021; Sovey et al., 2022). Notably, prior work did not always consider or discuss how social, cultural, and historical factors might have determined gendered inclinations toward CT, as exemplified by Wang et al. (2015), who showed how social encouragement is key for women pursuing careers in STEM and CS. While we acknowledge its importance, discussing sociocultural and historical determinants of CT is out of scope here.

1.3. Research questions

We extend prior work by inquiring about the factors predicting CT in constructionist game-based learning (GBL) through Multilevel Modeling (MLM) (Dedrick et al., 2009) and guided by the following research questions:

RQ1 *What factors, at group or individual level, predict CT in a constructionist GBL STEM curriculum?*

RQ2 *Do factors predicting CT interact?*

2. Methods

For our inquiry, we leverage a constructionist GBL STEM curriculum dataset focused on learning climate science, systems thinking, and CT in parallel (Puttick & Tucker-Raymond, 2018). As our inquiry accounts for students' gender, we do not stigmatize "pre-defined" gender roles in CS (Michell et al., 2018) by asserting that gender determines success in CT. Rather, we inquire how gender and other variables, including students' prior experience with *Scratch* and teachers' experience with the curriculum, might predict CT in interaction with hierarchical factors that shape CT learning. Compared to prior work (Kong et al., 2018), we explore factors predicting CT via MLM (Dedrick et al., 2009). Hence, to prior work showing one-dimensional, linear predictions between factors (e.g., thinking style, Durak and Saritepeci 2018) and CT, we integrate a comprehensive picture of how they are nested within the hierarchical structure of a learning environment, and to what extent the grouping factors within this hierarchy predict students' CT. The secondary dataset includes 483 games designed by pairs of 8th grade students in Scratch (Resnick et al., 2009) from 50 middle school science classes spread across five schools and taught by 11 teachers. Variables in this dataset include students' demographics (e.g., gender), student-designed game genre classification (e.g., *puzzle*, *simulation*), prior Scratch experience, and CT as scored by Dr. Scratch (Moreno-León et al., 2015). To compensate for missing data on gender and prior Scratch experience in the dataset, we perform data imputation (Song & Shepperd, 2007). Next, we describe (1) the STEM curriculum and its classroom activities, the student-designed games in Scratch, and their CT assessment via Dr. Scratch, and (2) data collection, data imputation techniques, and MLM analysis.

2.1. Constructionist game-based learning STEM curriculum for CT

The curriculum strongly emphasized artifact creation through Scratch and participatory pedagogy (Tucker-Raymond et al., 2019). The constructionist approach allowed students to actively shape their learning experience and engage in collaborative problem-solving through pair programming (Werner et al., 2004). This encouraged them to draw inspiration from their peers' creations and nurtured a sense of self-efficacy. Further, the resulting student-designed games and CT were assessed and monitored through Dr. Scratch (Moreno-León et al., 2015). Over three years, the curriculum was deployed in 35 middle school science classes spread across five distinct schools. This initiative involved the dedicated efforts of 11 teachers and saw an average of 20–30 students in each class, for an estimated 932 8th grade[2] students actively engaged in the curriculum. Notably, in our MLM, we consider only years two and three of the curriculum (i.e., *cohorts* 1 and 2), as year one was implemented by two teachers only and was considered a pilot.

2.1.1. Professional development, curriculum design, & classroom activities

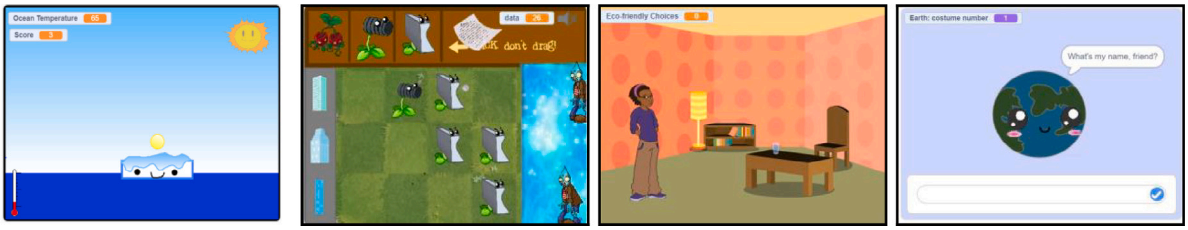
Teachers underwent professional development (PD) to familiarize themselves with Scratch and CT and integrate climate science and systems thinking. The teachers explored Scratch and its block-based programming with technology integration specialists and invested four hours experimenting with Scratch. The PD aligned with a participatory ethos (Rosebery & Puttick, 1997) and allowed teachers to (1) appreciate how systems thinking enriches students' comprehension of climate science and (2) identify strategies to tailor curriculum activities based on student's prior exposure to Scratch and attitudes towards CT. Nurturing distributed expertise (Cassidy et al., 2020) was facilitated by the curriculum. This approach was enhanced by online and physical materials covering climate change topics and principles underpinning computational problem-solving. The curriculum material also included instructional cards from Scratch, which facilitated the creation of coding routines and links to the Scratch community forums for addressing various queries. Also, the Creative Computing guide provided by *ScratchED*[3] allowed students to gather knowledge of scripting routines and access examples of code remixing. The curriculum encouraged creativity and cultivated a collaborative learning culture fostering students' peer critique, including pair programming and jigsaw grouping. Pair programming was supervised by teachers, who also helped sort students into pairs of either male–male (MM), female–female (FF), or male–female (MF) to balance gender mixing. Also, teachers publicly identified students with prior Scratch expertise to help others when asked, promoting remixing publicly accessible Scratch projects to practice coding (Kafai & Burke, 2017). Finally, the *Triadic Game Design* (TGD) framework for serious game design (Harteveld, 2011) was employed to help students articulate reality, meaning, and play in their gamified adaptation of climate science.

Students and teachers explored climate change, such as global warming and energy consumption. With the supervision of teachers, the students selected a climate change topic to represent in their Scratch projects. For inspiration, students played existing video games on climate change from NASA's *ClimateKids*.[4] Students critically analyzed the design elements of these games that convey educational messages to discern the distinctive characteristics of serious games. They participated in a 10-block challenge following the *ScratchED* guide to familiarize themselves with Scratch by programming a simple application using only ten blocks. After the 10-block challenge, the students collaboratively refined their ideas for climate science games and worked on their Scratch projects in pairs.

[2] In this paper, we adopt the US conventional school grade system (e.g., 5th grade and beyond) when referencing different levels of learning.

[3] https://scratched.gse.harvard.edu/index.html

[4] https://climatekids.nasa.gov/menu/play/



(a) Albedo Pong - *Pong* (b) Plant vs Zomseas - *Clicker* (c) Eco-Friendly - *Quiz* (d) Pet Earth! - *Simulation*

Fig. 1. Examples of student-designed games about climate change in Scratch, title, and game genre (in *Italics*).

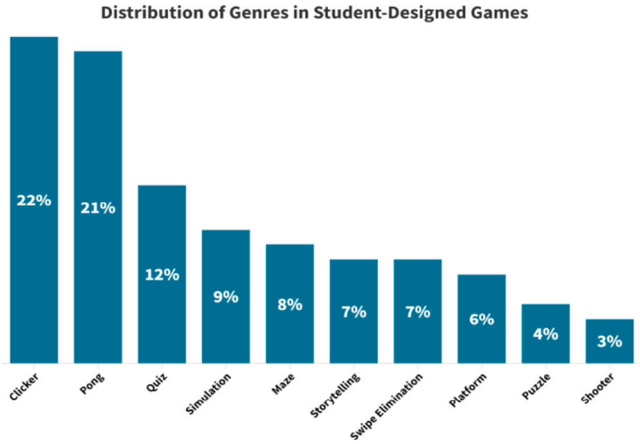


Fig. 2. The distribution of game genres for student-designed games, from more (*clicker*) to less (*shooter)* popular.

2.1.2. Student-designed games

The students assigned distinct tags/names to their projects and did not include their real names to comply with IRB requirements. The games were saved in an external repository, enabling us to access the data for analysis. We leveraged the Scratch developers' API to capture and gather information from the compressed .sb2 files, including block usage and timestamps. The student-designed game genres were previously coded through emergent and a priori codes using TGD and Heintz and Law's categorization of game genres (Heintz & Law, 2015). The game genres were assigned a *primary* and *secondary* game genres, as *action* games were prominent in the dataset (68%) compared to the rest (32%). Such a breakdown of *action* genre into *secondary* game genres allowed for balance within our MLM analysis. The original coding underwent several iterations across the entire dataset until refined coding. Fig. 1 shows four examples of student-designed games and their game genre. Among the 483 games we collected, 102 games were either unplayable, had less than ten games within a game genre (i.e., *adventure* and *strategy*), or were missing from the repository. Fig. 2 shows the genre distribution of the 381 student-designed games included in the MLM.

2.1.3. Assessing CT via automated metrics

Automated metrics can help monitor students' progress in CT (Troiano et al., 2019), give scholars a rubric that can systematically score CT (Moreno-León et al., 2015), and provide visual feedback to examine CT skill development at a glance (Vieira, 2020). We assessed student-designed using Dr. Scratch (Moreno-León et al., 2015), which was previously used to assess CT in game-based STEM curricula (Troiano et al., 2020a). Dr. Scratch quantifies CT within Scratch projects on a scale from 0 to 3 concerning seven designated CT dimensions (Table 1). Each assigned score defines a level of CT proficiency: 0 *null*, 1 *basic CT*, 2 *developing CT*, and 3 *CT proficiency*. The cumulative score is computed as the sum of individual scores across each CT dimension, yielding a maximum of 21. These scores are based on observed coding practices within Scratch. For instance, Dr. Scratch assigns 1 point for using basic logic (e.g., *if* blocks), while 3 points are assigned when using more complex logic operators (e.g., *AND/OR*).

2.2. Data collection

For our MLM, we use a secondary dataset aggregating the curriculum data (i.e., student-designed games and CT scores as assessed by Dr. Scratch) with survey data completed by students and their teachers. School-district-based technology specialists and curriculum teachers designed the curriculum survey. The survey gathered students' (1) race/ethnicity, (2) gender, (3) preferred game to play, (4) prior experience with Scratch and game design, (5) average weekly hours for internet usage, (6) average weekly hours spent playing video games, and (7) technology use. The survey combined dichotomous scales (i.e., *yes/no*), multiple and single

Table 1
Dr. Scratch metric scale, showing competence level for each CT dimension and relative Scratch practices.

Term	CT Dimension		Competence level		
	Definition	Null (0)	Basic (1)	Developing (2)	Proficient (3)
Abstraction	The ability to conceptualize and then represent an idea or a process in more general terms (Weintrop et al., 2016)	–	More than one script and more than one sprite	Definition of blocks	Use of clones
Data representation	Representing data through abstractions, such as models and simulations (Barr & Stephenson, 2011)	–	Modifiers of sprite properties	Operations on variables	Operations on lists
Flow control	A high-level way of programming a computer to make decisions, simple or complicated, executed once or multiple times (Scopatz & Huff, 2015)	–	Sequence of blocks	Repeat, forever	Repeat until
Logic	Conditionals and rules that allow building up and representing complex ideas (Scopatz & Huff, 2015)	–	If	If else	Logic operations
Parallelism	Handling multiple scripts or sequences of code that run simultaneously (Park & Shin, 2019)	–	Two scripts on green flag	Two scripts on key pressed, two scripts on sprite clicked on the same sprite	Two scripts on when I receive message, create clone, two scripts when %s is >%s, two scripts on when backdrop change to
Synchronization	The coordination of simultaneous threads or processes	–	Wait	Broadcast, when I receive message, stop all, stop program, stop programs sprite	Wait until, when backdrop change to, broadcast and wait
User interactivity	Designing and programming for user input	–	Green flag	Key pressed, sprite clicked, ask and wait, mouse blocks	When %s is >%s, video, audio

answers, and open box answers (see Appendix A). Previous experience with Scratch used a 5-point Likert scale, where each point of the scale was matched to a statement formulated by teachers that students could easily understand. Scratch experience refers to the experience students had with Scratch programming before the curriculum. They self-assessed their experience on a scale from less (Huh?) to more experienced (A lot of experience, see Table 3, *Class* column). Notably, students' prior experience with Scratch was not gathered through the survey but by teachers during pair formation and later added to the survey results (see Appendix B). Such an "ad hoc" form of surveying (or questioning) is common practice in children and education research (Greig et al., 2007), and while it may hinder replicability, it is often needed to fit inquiries to specific contexts.

We wrote software scripts to assist with combining the information related to the curriculum data and the student demographics into one dataset. Since the games were stored on the Scratch website, information about the project name, student names, and IDs could be retrieved by running short JavaScript snippets. These snippets could read the HTML code and automatically gather information. Game scores were assessed by Dr. Scratch, and we used the student-designed game ID as a key and assigned the computed scores to the student pairs whose IDs matched the project ID. Two researchers and a data analyst performed, inspected, and verified this process.

2.2.1. Data imputation of missing data

In the curriculum dataset, missing data comprised gender and Scratch experience, which are key variables to be considered for our analysis. Of 483 projects to be analyzed, 18.7% gender data and 43.3% prior Scratch experience data were missing (Table 2). Notably, participation in the curriculum was voluntary. Hence, factors like absenteeism (a common issue in education research, Cheema 2014) may have contributed to missing data. To address the missing data, we performed data imputation.

We searched for models that could reliably infer gender based on the students' names and leveraged the work of Karimi et al. (2016) using the R package provided in their paper[5] and Gender API.[6] The API was ranked first in performance and error rate in several studies where gender data was absent (Prana et al., 2021; Santamaría & Mihaljević, 2018; Sebo, 2021) and uses database lookup and classification algorithms that proved effective for missing data retrieval. While name-based gender retrieval models are far from perfect and are being scrutinized for their potential biases (Mihaljević et al., 2019), Karimi et al.'s Gender API is trusted by the scientific community, as it incorporates a retrieval method that *"takes into account how gendered naming practices have changed over time"* (Blevins & Mullen, 2015, p. 24). The sources used to create the package's dataset are derived from the U.S. Social Security

[5] https://www.rdocumentation.org/packages/gender/versions/0.6.0
[6] https://gender-api.com/en/about-us

Table 2
The predicted prior experience with Scratch of students vs their predicted gender, along with the total count of their projects (also expressed in percentage %).

Gender_Predicted	Experience_Predicted		Total
	N	Y	
N	237 (49.0%)	156 (32.3%)	**393 (81.3%)**
Y	37 (7.7%)	53 (11%)	**90 (18.7%)**
Total	**274 (56.7%)**	**209 (43.3%)**	**483 (100%)**

Table 3
Precision and recall table for the five levels of prior experience with Scratch.

Class	Precision	Recall
Huh?	76%	78%
I have heard of it	79%	83%
Hour of code	79%	74%
Created a project	75%	70%
A lot of experience	64%	67%

Administration, the U.S. Census Bureau, and the North Atlantic Population Project. While limited to American names, the sources did not hinder our gender inference as the students involved in the curriculum were all based in the U.S.

We initially queried students' genders using the API V1.0, retrieving the majority of missing genders. We then used the V2.0 endpoint to fix the remaining missing genders. We first validated the Gender API with a sample of 820 students whose gender information was already known, and the API correctly classified 96.7% of the cases (793 out of 820 students). For the 207 students with missing gender data, the Gender API provided results with a confidence level of 70% or higher in 93.7% of cases. For missing data on prior Scratch experience, which affected 43.3% of the dataset, we used data imputation with classifiers specifically designed to predict missing values. Among the models tested—(1) *softmax classifier* (Wang et al., 2019), (2) *decision tree* (Nikfalazar et al., 2020), and (3) *random forest* (Tang & Ishwaran, 2017)—the *random forest* method provided the most robust results. With 80 estimators and stratified K-fold cross-validation ($k = 5$, Camacho and Ferrer 2012), the model achieved an accuracy of 76.8%, as shown in Table 3, and a "Receiver Operating Characteristics - Area Under the Curve" (ROC-AUC) score of 92.1%. This approach ensured that the imputation process was reliable, aligned with observed data patterns, and mitigated the risk of overfitting.

While imputing 43.3% of the Scratch experience data may raise concerns, excluding such a large proportion of the data would have significantly reduced the representativeness of our analysis. Our imputation strategy followed best practices in education research (Cheema, 2014), and the random forest model's superior performance provides a reliable foundation for this approach. To validate the inclusion of imputed data, we conducted parallel analyses of our MLM models with and without missing data. As shown in Appendix C, the results were consistent across both approaches, indicating that imputing the data did not alter the findings' overall trends or statistical significance. By including imputed data, we ensured that a larger portion of the dataset could be utilized, thereby increasing the statistical power and robustness of the modeling. Importantly, this decision did not affect the interpretation of key results. We acknowledge that imputation introduces assumptions about missing data, and we encourage future work to replicate and extend our findings by exploring alternative approaches.

2.3. Multilevel modeling (MLM) of factors predicting CT

MLM is a statistical method for analyzing complex hierarchical data (Gelman, 2006) and is suited for analyzing nested data structures, such as the context of students within classrooms or schools. MLM also extends linear regression to analyze hierarchical structures implicit in a dataset (Hox & Roberts, 2011). As such, it effectively captures the hierarchy of factors interacting with CT by modeling individual and group-level variations, either fixed or random effects. Compared to Structural Equation Modeling (SEM, Kong et al. 2018), MLM better aligns with the hierarchical configuration inherent in the dataset and our need to inquire about individual and group-level predicting factors. By incorporating group-level (e.g., teacher) and individual-level (e.g., game genre) attributes, MLM effectively captures shared attributes among observations within the same group. In that respect, it allows us to (1) understand how context may affect individual outcomes (here, the CT score of student-designed games) and (2) identify subgroups within the data with different relationships between predicting factors and the CT score (Leyland & Groenewegen, 2020).

We start by identifying what group-level attributes significantly predict the CT score of student-designed games based on the natural hierarchy in the dataset, namely (1) individual observations (i.e., student games), (2) cohort, (3) teacher, and (4) school (see Fig. 3). Once we identified levels that significantly predict CT scores at the group level, we looked for factors that are fixed effects (i.e., independent variables) and may predict CT at the individual level—these include (1) students' gender (here, gender pairing as MM, FF, or MF), (2) student's race/ethnicity, (3) teacher experience implementing a game-based STEM curriculum, (4) student prior experience with Scratch, (5) game genre of student-designed games, (6) average weekly hours spent on playing video games (7) average weekly hours of internet usage, (8) prior experience in game development, (9) use of technology, and (10) preferred game to play at home.

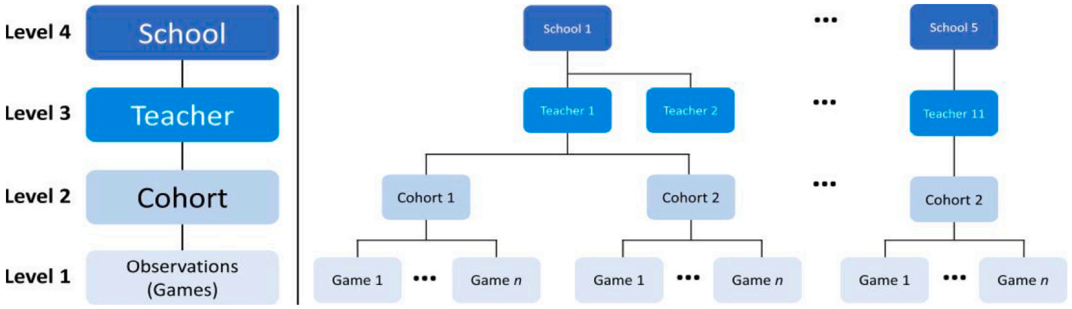


Fig. 3. The hierarchical structure in the dataset analyzed through MLM (left) and the nested structure in our MLM (right); the number *n* represents the end number of games in a cohort and varies based on teacher/cohort.

Table 4
P-value and ICC for potential attributes in grouping observations (levels).

Independent variable	P value	ICC
School	1	0.00
Teacher	**1.4e−08** ***	**0.10**
Cohort	0.245	0.01

*** Significant at $p < 0.001$.

2.3.1. MLM implementation and analysis

The hierarchical MLM structure we test for the significance of predicting CT scores is depicted in Fig. 3. The lowest level comprises each observation (i.e., a student game), which can be hierarchically nested within *cohort*, *teacher*, or *school.* Among the schools analyzed, two schools involved four teachers each, while the remaining three schools had one teacher each, totaling five schools and 11 teachers across two cohorts. Notably, these schools and teachers did not participate uniformly across all cohorts. Specifically, four teachers in two schools were actively involved in both cohorts (i.e., deemed as *experienced* teachers in *cohort 2* and *new* in *cohort 1*). By contrast, an additional seven teachers contributed only to *cohort 2* (i.e., deemed as *new*). Consequently, we observe a total of 15 distinct teacher–cohort combinations. For our MLM, the dependent variable is the aggregated CT score (ranging from 0 to 21) as assessed by Dr. Scratch. Independent variables include the 10 mentioned at the end of Section 2.3. As recommended by prior work with MLM (Hox et al., 2017; Hox & Roberts, 2011), we start by (1) constructing a baseline model including only significant levels, then (2) incorporating significant independent variables, and finally (3) assessing interaction effects among independent variables.

Model 0: Baseline—We followed Hox et al. (2017, 2011) and constructed MLM models incrementally. We started by building a random intercept model with only the cohorts at the second level to group the observations (i.e., student games). We added further levels to the model and retained only those demonstrating statistical significance. In this phase, we focused on identifying potential levels (i.e., school, teacher, cohort) that show statistical significance at the group level. After evaluating all potential grouping variables for significance (Table 4), we found that only *teacher* was statistically significant (p = 1.4e−08) and included it in the next models. Notably, none of the models with more than two levels exhibited statistical significance (all having p >.05). Consequently, we devised our baseline model, denoted as *Model 0*, constituting a multilevel framework comprising two levels: (1) student games at the first level and (2) teachers as the grouping factor at the second level.

Model 1: Intermediate—Keeping teachers and games as distinct levels (same as *Model 0*), we incorporated all candidate independent variables and expanded the model. Following recommendations from Hox, since cohort and school didn't prove significant in *Model 0*, we added them at the individual level along with the other 10 independent variables mentioned in Section 2.3. We focused on determining whether any independent variables significantly predict CT scores during this stage. We retained only the statistically significant independent variables, employing a backward elimination approach, leading to a refined *Model 1*. This streamlined model encompasses (1) game genre, (2) students' prior experience with Scratch, (3) teachers' experience implementing the constructionist GBL curriculum, and (4) cohort as predictors of CT, with *teacher* as a significant grouping level. This approach enabled us to validate the model's capacity to incorporate hierarchical levels and independent variables.

Model 2: Final—The difference between *Model 1* and *Model 2* is that *Model 1* only considered independent variables without their interactions. We generated extended versions of *Model 1* that incorporate interactions between different independent variables, evaluating the statistical significance of six possible interactions:

- Students' prior experience with Scratch<—>Teachers' experience implementing the curriculum
- Students' prior experience with Scratch<—>Student-designed game genres
- Students' prior experience with Scratch<—>Cohort
- Teachers' experience implementing the curriculum<—>Student-designed game genres
- Teachers' experience implementing the curriculum<—>Cohort
- Student-designed game genres<—>Cohort

Table 5
The three MLM models and their statistical comparison.

Model	npar	AIC	BIC	logLik	Chisq	Df	Pr(>Chisq)
Model 0	3	2288.7	2301.3	−1141.4			
Model 1	19	2208.9	2288.3	−1085.4	111.871	16	**2e-16** ***
Model 2	29	2208.4	2329.7	−1075.2	20.417	10	**0.02555** *

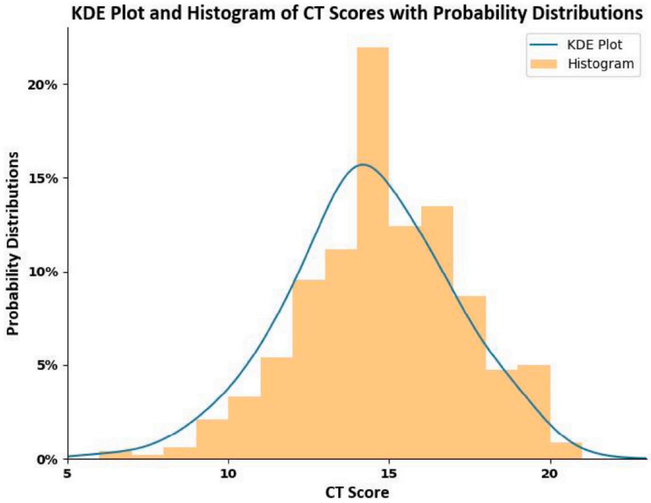


Fig. 4. Actual distribution and estimated normal distribution (KDE plot) of CT scores for student-designed games.

We tested the correlation between these interactions and CT scores, finding only the interaction between teacher experience and game genre to be statistically significant (p = .026). Consequently, *Model 2*, which incorporates this interaction into *Model 1*, was selected as the final model, capturing the significant correlation of grouping levels and independent variables with CT score while accounting for significant interactions (see Table 6).

2.3.2. MLM models comparison

Table 5 compares the three models from our MLM. Notably, *Model 1* and *Model 2* show a better fit to the data than our baseline *Model 0*, with p-values less than .05. We also consider Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC). While the AIC imposes a linear penalty on the number of variables in the models, the BIC applies a logarithmic penalty (Hox et al., 2017; James et al., 2013). These metrics assess how well the models fit the data while considering their complexity. *Model 1* and *Model 2* are more complex than *Model 0* as they include more variables (19 and 29, respectively; only 3 for *Model 0*). Hence, through AIC and BIC, we evaluate if the improved fit is worth the increased complexity. As complexity is integrated into the model, the AIC decreases (from 2288.7 to 2208.9 to 2208.4). Conversely, the BIC initially declines from *Model 0* to *Model 1* (from 2301.3 to 2288.3) but subsequently rises due to a more stringent penalty (reaching 2329.7). Through a comprehensive fitness evaluation, along with AIC and BIC, we find merit in *Model 1* and *Model 2*. By incorporating independent variables and their interactions, these models offer valuable insights into the factors predicting CT while addressing overfitting. Further, following James et al. (2013), we assessed the absence of multicollinearity among variables of *Model 1* and *Model 2* by employing a variance inflation factor (VIF) metric. The requirement was that each variable have a VIF value of less than 5. In *Model 1*, the maximum VIF value was 1.63; in *Model 2*, it reached 3.21, both linked to teacher experience. This confirms the absence of multicollinearity, as all VIFs in both models are below 5.

3. Results

Among 932 students, 440 (47%) were female, and 492 (53%) were male. Fig. 5 shows the distribution of gender pairing characterizing pair programming, while Fig. 4 shows the distribution of CT scores for student-designed games. We show the distribution of scores (yellow bars) and an estimated normal distribution (blue line) through a Kernel Density Estimation (KDE) plot in Fig. 4. We integrate a smooth curve overlaying the histogram to approximate the probability density function. By juxtaposing these two distributions, we highlight central tendencies (M = 14.33) and variations (SD = 2.59) in CT score. Finally, Fig. 6 shows the distribution of student-designed games organized by gender pairs. Here, design preferences within specific game genres are evident, where FF designed *quiz* and *clicker* games frequently. By contrast, MM designed more *pong* games.

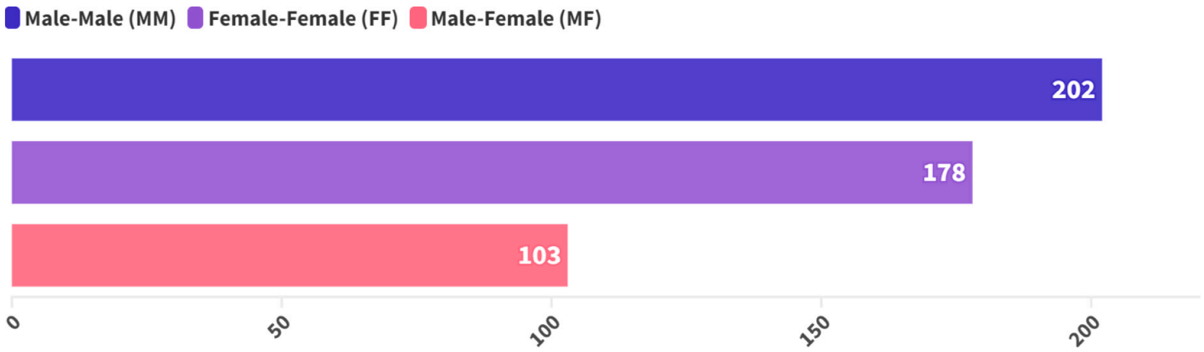


Fig. 5. The distribution of gender pairings across the two cohorts in analysis.

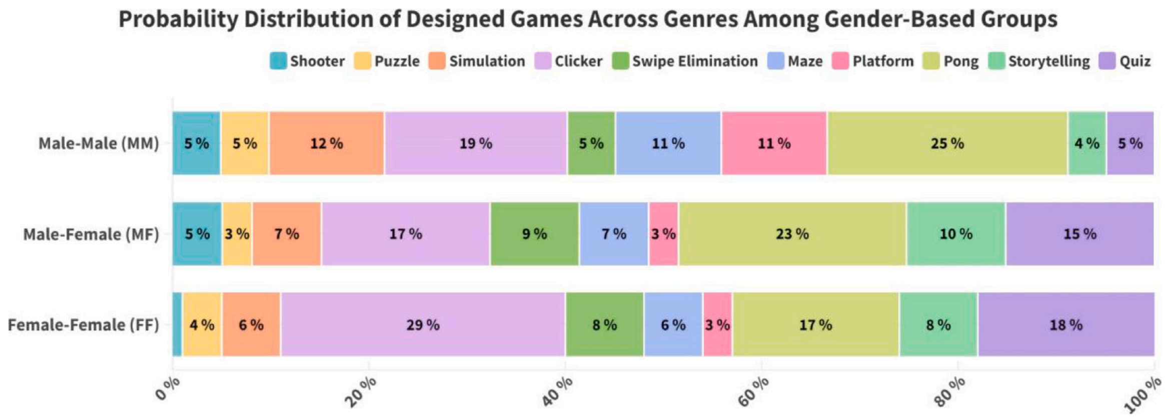


Fig. 6. The distribution of student-designed games based on gender pairs and game genres.

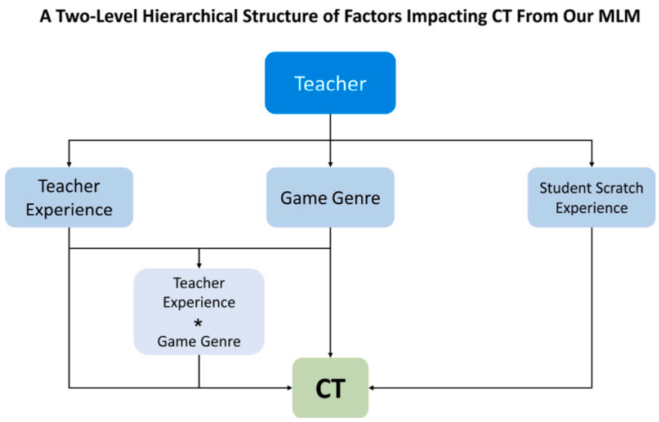

A Two-Level Hierarchical Structure of Factors Impacting CT From Our MLM
Teacher
Teacher Experience
Game Genre
Student Scratch Experience
Teacher Experience * Game Genre
CT

Fig. 7. The MLM structure of factors predicting CT.

3.1. *Predicting factors and their hierarchy (RQ1)*

Fig. 7 shows the hierarchical structure of predicting factors, highlighting how *teacher* predicts CT at the group level and mediates other significant factors at the individual level. Table 6 shows the resulting MLM models, fixed effects (independent variables), random effects (*teacher* as grouping variable), and statistical significance metrics (p-values) for each variable and model. We identified four main predictors: (1) *teachers and their experience implementing the constructionist GBL curriculum*; (2) *students' prior Scratch experience*; (3) *student-designed game genre*; and (4) *interaction between teacher experience and game genre* with predictions by *teachers* at the group level (Fig. 7).

Table 6
The resulting MLM models, fixed effects (independent variables), and statistical values.

Fixed effects	Model 0: *Baseline*		Model 1: *Intermediate*		Model 2: *Final*	
	Estimate	P-value	Estimate	P-value	Estimate	P-value
Intercept	14.26	**2.21e−14** ***	16.53	**3.16e−06** ***	14.83	**1.74e−05** ***
Main Effects						
Experience-2			0.17	0.563	0.21	0.457
Experience-3			0.45	0.145	0.37	0.224
Experience-4			1.69	**3.32e−06** ***	1.47	**4.82e−05** ***
Experience-5			2.11	**5.80e−06** ***	1.91	**3.11e−05** ***
Teacher Experience-New			−0.42	0.345	2.21	**0.011** *
Genre-Clicker			0.85	**0.027** *	2.85	**0.0003** ***
Genre-Maze			0.74	0.151	3.00	**0.0021** **
Genre-Platform			0.61	0.275	2.94	**0.001** ***
Genre-Pong			0.20	0.614	2.53	**0.0005** ***
Genre-Puzzle			1.86	**0.004** **	4.66	**6.12e−06** ***
Genre-Quiz			−1.13	**0.015** *	1.08	0.225
Genre-Shooter			1.90	**0.007** **	3.43	**0.007** **
Genre-Simulation			1.27	**0.013** *	3.81	**8.28e−06** ***
Genre-Storytelling			−0.18	0.744	−0.06	0.973
Genre-Swipe elimination			0.28	0.601	2.46	**0.005** **
Cohort-2			−1.00	**0.044** *	−1.09	0.069
Interaction Effects						
Teacher Experience-New: Genre-Clicker					−2.73	**0.002** **
Teacher Experience-New: Genre-Maze					−3.13	**0.006** **
Teacher Experience-New: Genre-Platform					−3.33	**0.005** **
Teacher Experience-New: Genre-Pong					−3.35	**0.0001** ***
Teacher Experience-New: Genre-Puzzle					−4.10	**0.002** **
Teacher Experience-New: Genre-Quiz					−2.90	**0.005** **
Teacher Experience-New: Genre-Shooter					−1.93	0.201
Teacher Experience-New: Genre-Simulation					−3.71	**0.0007** ***
Teacher Experience-New: Genre-Storytelling					−0.73	0.685
Teacher Experience-New: Genre-Swipe elimination					−3.15	**0.005** **
	Model 0		Model 1		Model 2	
Random Effects	**logLik**	**P-value**	**logLik**	**P-value**	**logLik**	**P-value**
Level 2: Teacher	−1144.2	**1.43e−08** ***	−1088.0	**0.025** *	−1081.1	**0.0006** ***

3.1.1. Teachers and their experience with curriculum implementation

Model 0 is based on a two-level structure. The first level considers individual observations (i.e., student-designed games). The second level groups these observations based on *teacher*. We have listed the attributes used for grouping games at this second level in Table 4. When applying MLM to this structure, we found that the teacher has a significant influence on how games are grouped among the attributes we consider at the second level ($p = 1.4 \times 10^{-8}$ as indicated in Table 4). Simply put, different teachers guide students' game design differently. We measure this through Intraclass Correlation Coefficient (ICC).[7] ICC tells us how similar or different the games are within the same teacher's group. ICC for *teacher* level indicated that the teacher's influence on game grouping is substantial (ICC = 0.10). When introducing *school* as another level in the model, we encountered challenges as the model struggled to converge, showing how *school* was not as effective as *teacher* as a level for grouping.

Fig. 8 highlights the contrast between *experienced* teachers (in green) and *new* teachers (in orange, limited to *cohort 2*). The data reveals that most teacher groups exhibit normally distributed CT scores. On average, *experienced* teachers demonstrate higher CT scores compared to *new* teachers. However, an exception is observed with T-5 (a new teacher), who achieves the highest average CT score (15.45). In contrast, T-7 (also a new teacher) records the lowest average score (12.26), with a variability of 1.23 SD observed across the average Dr. Scratch scores for all teacher groups. We analyzed score distribution across cohorts, revealing a notable decrease (p = .04, based on *Model 1* in Table 6 and Fig. 9) in CT scores within *cohort 2* when compared to *cohort 1*. In *Model 1*, teacher experience was not statistically significant (p = .35, Table 6). However, it becomes significant when including interactions in *Model 2* (p = .01, Table 6). The MLM shows that the significant prediction of CT score using cohort in *Model 1* is a result of its correlation with teacher experience. Furthermore, this is accentuated by the disparate genre distributions observed within *cohort 1* and *cohort 2*. We then compared the scores of *experienced* and *new* teachers within *cohort 2* to look for further clarification and unfold this complexity. We found teacher experience predicting CT, showing that games supervised by *experienced* teachers yielded higher CT scores than *new* teachers (Fig. 9).

[7] ICC is used in MLM to quantify the similarity among units of the same group. It applies to data organized into groups (see Hox et al. 2017).

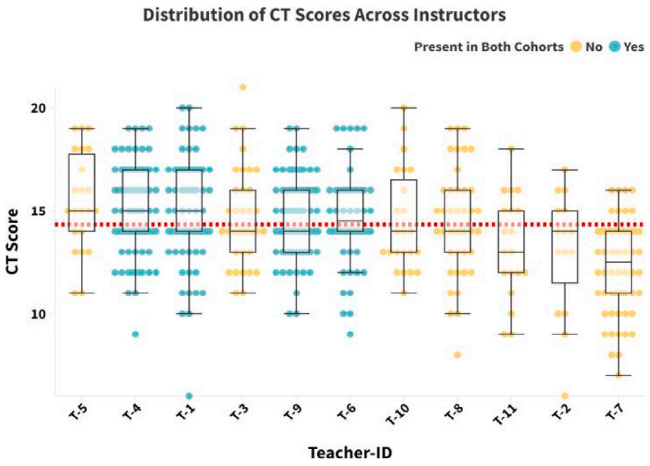


Fig. 8. Significant difference in CT scores across different groups (teachers): T-5 shows the highest average score of 15.45, while T-7 shows the lowest average of 12.26.

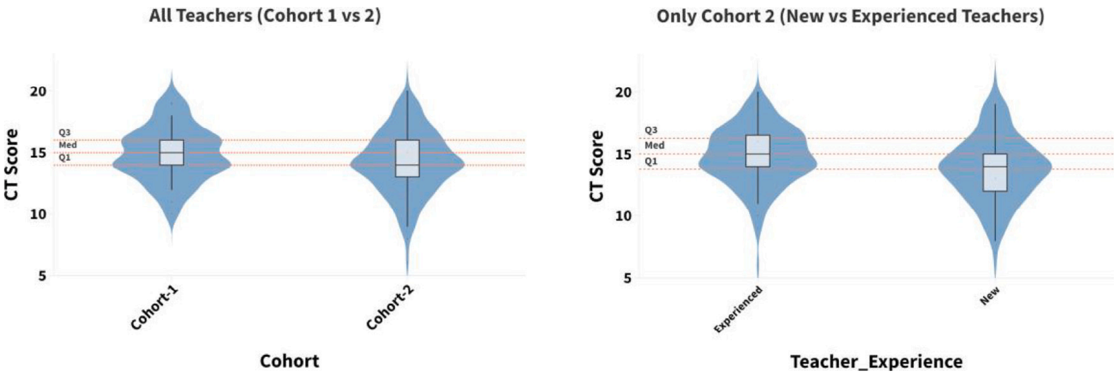


Fig. 9. Comparing CT scores between *cohort 1* and *cohort 2* (left), showing a significant decrease in CT scores for cohort 2 (p = .04, *Model 1* in Table 6). Within *cohort 2* only (right), students supervised by *experienced* teachers achieved a higher average CT score, consistent with our MLM results that incorporate interaction effects (*Model 2* in Table 6).

3.1.2. Student prior experience with scratch

Fig. 10 shows a red dotted line representing the average CT score of students. Students with prior Scratch experience at levels 4 and 5 significantly have higher CT scores (p <.001), while students at levels 1, 2, and 3 do not significantly deviate from the average CT score and tend to have similar CT performances.

3.1.3. Game genre

We found that game genre predicts CT, consistent with prior HCI work (Troiano et al., 2020a). Further, we found discrepancies in CT scores across game genres, as depicted in Fig. 11 and detailed in Table 6. This trend is observable in *Model 1*. Among game genres, *shooter*, *puzzle*, *simulation*, and *clicker* stand out with their higher CT score compared to other genres; *quiz* has the lowest score. Notably, the correlation of these game genres with CT scores is statistically significant, as highlighted in green in Fig. 11.

3.2. The interaction between teacher experience and game genre (RQ2)

We found that *Model 2* incorporating interaction between teacher experience and game genre proved statistically significant (p = .026). Notably, *shooter* and *storytelling* did not interact significantly with teacher experience (p = .20 and .69, respectively). This interaction is observable in Fig. 13, where game genres predict CT scores in conjunction with teaching experience. The interaction effect underscores that *experienced* and *new* teachers yield diverse effects on CT scores across game genres. Specifically, *puzzle* stands out among the genres with significant interaction, showing the highest average difference of CT scores between *experienced* and *new* teachers. In *puzzle*, games crafted under the guidance of *experienced* teachers garnered +1.88 CT points on average compared to *new* teachers. Similarly, *simulation* and *pong* show notable differences (i.e., +1.50 and +1.14 in average CT scores for *experienced* teachers, respectively). Instead, on average *new* teachers showed better performance in *shooter* (+0.28 CT points) and *storytelling* (+1.48 CT points); this, however, did not show statistical significance (p = .20, and .69 respectively, Fig. 13 and Table 6). The relationship

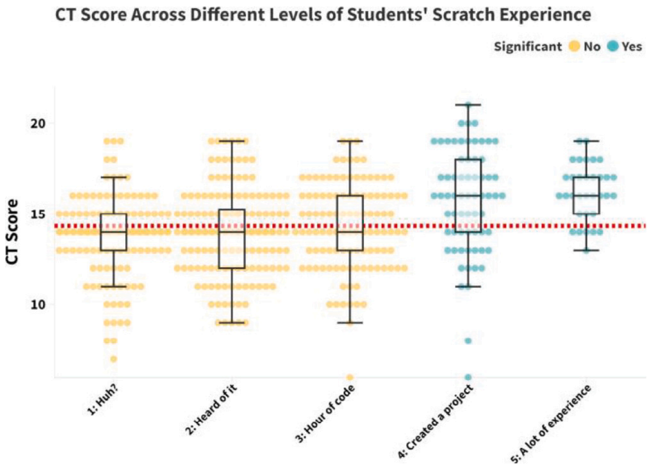


Fig. 10. The distribution of CT score across different levels of students' prior experience with Scratch.

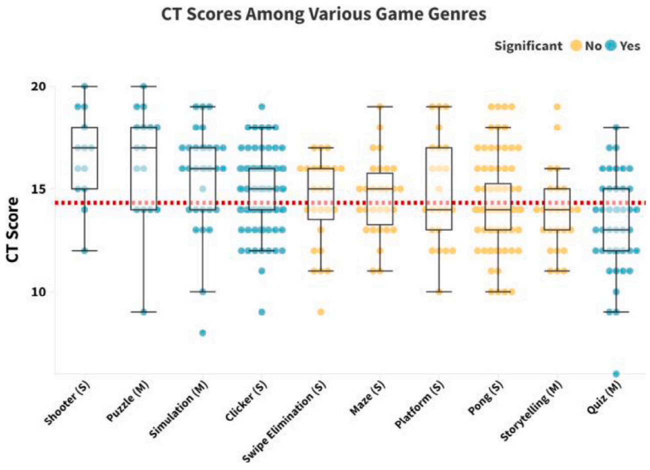


Fig. 11. Significant difference in CT score across student games: *shooter* and *puzzle* had the highest median CT score (*Median* = 17), while *quiz* had the lowest (*Median* = 13). (M) = primary game genres; (S) = sub-genres of *action*.

between the presence of *experienced* teachers in distinct genres and the higher CT scores for games overseen by *experienced* teachers becomes more evident when combining Figs. 13 and 12.

These visualizations highlight the statistical relationship between teacher experience and CT scores, as well as differences in genre selection among students supervised by *experienced* versus *new* teachers. However, the mechanisms driving these differences remain unclear and are beyond the scope of this quantitative analysis, as we later elaborate in Sections 4.2 and 4.3. We organized the genres in Fig. 12 according to the same sequence seen in Fig. 13, with descending order based on the difference of CT score for *experienced* and *new* teachers. Fig. 12 reveals that the first five genres exhibit a higher selection probability among experienced teachers, while the subsequent five genres show a higher selection probability among new teachers.

4. Discussion

In response to our initial research questions, the results of our MLM lead to the following insights:

RQ1 *What factors, at group or individual level, predict CT in a constructionist GBL STEM curriculum?*

- *Group Factors:* within the implicit hierarchy in the secondary dataset under analysis, teachers emerged as a significant predictor, while cohort and school did not;
- *Individual Factors:* with teachers as the grouping factor, (1) teacher experience with implementing a GBL STEM curriculum, (2) students' prior experience with Scratch, and (3) the game genre in student-designed games emerged as significant predictors of CT uptake.

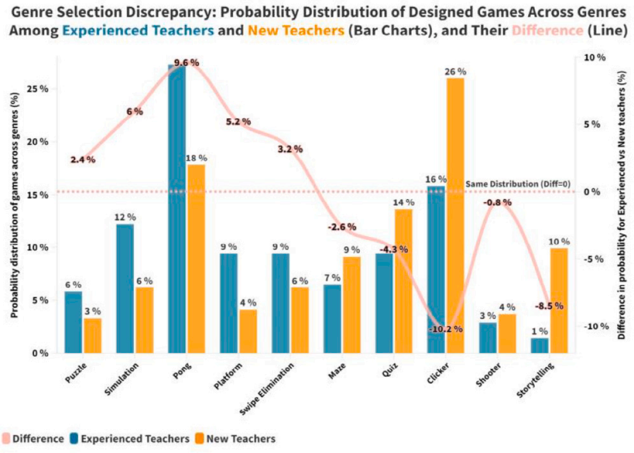


Fig. 12. The variability in game genre preferences among *experienced* and *new* teachers. Notably, *puzzle*, *simulation*, *pong*, *platform*, and *swipe elimination*, are favored by *experienced* teachers, who exhibit a positive probability difference (pink line). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

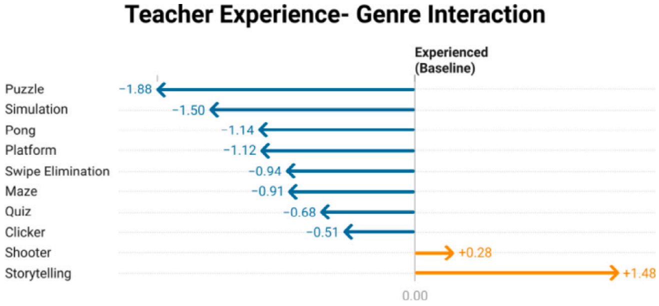


Fig. 13. CT score tendencies across all game genres based on *teacher* and using *experienced* teachers as a baseline. Notably, *new* teachers performed better in *shooter* and *storytelling*, although statistical analysis did not show significance.

RQ2 *Do factors predicting CT interact?*

- Our MLM shows that teacher experience and student-designed game genres interact significantly in predicting students' CT.

4.1. Lessons learned from using MLM to assess CT in constructionist GBL

While MLM is widely used in education, psychology, and social science research (Dedrick et al., 2009), it was not used before to inquire about factors predicting students' CT in constructionist GBL. The hierarchical relationships and interactions predicting CT emerging from our MLM meaningfully extend prior research inquiring about factors predicting CT through SEM (Durak & Saritepeci, 2018), multiple regressions (Relkin et al., 2021), and multinomial logistic regression (Atman Uslu, 2023). Compared to these prior inquiries, our MLM revealed the structural and hierarchical nuances of the relationships among factors predicting CT, providing novel insights into what actors (e.g., teachers) and features (e.g., game genres) may impact CT development in a constructionist GBL curriculum. In that respect, MLM allows for effective and appropriate modeling of these relationships within the nested data structure, while avoiding potential statistical errors from violating independence assumptions, which may be the case with statistical models used in prior similar work, such as multiple linear regressions (Relkin et al., 2021) or multinomial logistic regression (Atman Uslu, 2023; Hox & Roberts, 2011). In particular, MLM allowed us to partition the variance, highlighting the role of significant individual-level predictors of CT in constructionist GBL that teachers mediate. Below, we discuss the mediating role of teachers and the role of game genres, which is the individual-level predictor with the most significant role out of the individual-level predictors.

4.2. Role of game genres

Our MLM revealed the structural and hierarchical nuances of the relationships among factors predicting CT compared to prior work. Particularly, we showed how the game genre of student-designed games was a prominent predictor (see Figs. 7 and 11). Considering the strong emphasis on game design in the GBL curriculum, certain outcomes, such as the significance of game genre, are somewhat anticipated and align with earlier studies (Troiano et al., 2020a). However, we extend these studies by revealing a statistically significant interaction between teacher experience and game genre (highlighted in Fig. 13 and Table 6), showing their potential combined impact on CT uptake. We provide quantitative evidence (highlighted in Fig. 12) to shed more light on the interaction between teacher experience and game genre, indicating that experienced teachers may subtly influence students' choices, potentially favoring genres with higher CT demands, like puzzles and simulations. However, the mechanisms underpinning this relationship remain unclear and are beyond the scope of our quantitative analysis. Future research may use classroom observations and teacher interviews to explore how pedagogical strategies shape GBL and CT uptake—for example, understanding how and if experienced teachers steer students toward certain genres, and how they guide students differently compared to less experienced teachers will meaningfully extend the interpretation of our results.

Furthermore, whether these predicting factors will change based on curriculum structure and core learning activities, either game-based or non-game design curricula, should be further inquired. For instance, consider CT uptake through educational robotics (Angeli & Valanides, 2020), unplugged activities (Rijke et al., 2018), or music (Freeman et al., 2019)—what would be the relative of game genre? While we do not have an answer, we encourage researchers to use MLM across diverse CT curricula to unpack hierarchical structures and extend an understanding of what factors predict CT. While we evidenced a possible influence of game genre on CT scores, establishing causation is challenging. Simple click-based games may demand less complex logic than more intricate genres like shooter games. Hence, it is reasonable that certain game genres would "naturally boost" CT as they require advanced programming and CT skills. Our final MLM model provides valuable insights into this matter. The analysis revealed that the interaction between students' prior Scratch experience and game genres lacked statistical significance. In short, no discernible pattern emerged, suggesting that students with higher prior Scratch experience, a potential proxy for advanced CT skills, did not choose to design games that potentially lead to high CT scores. Additionally, we did not find significance in CT score differences across genre-experience combinations. Future work should investigate the relationship between students' CT and game complexity to help explain further the relationship between game genre and CT score.

4.3. Role of teachers

Our MLM further shows that teacher experience implementing a game-based curriculum for CT uptake predicts students' CT performance. These results resonate with prior work on CT uptake through educational robotics (Angeli & Valanides, 2020) and game-based learning (Jin et al., 2021). To understand the implications of our MLM results, we must consider the correlation of game genre with CT uptake contingent on teachers' prior experience implementing the curriculum—this is particularly relevant to educators who engage with the adoption of (1) "new" learning technologies like Scratch and (2) implement constructionism (Papert, 1980). Prior work on educational pedagogy and 21st century teaching argued that *"educators need to be knowledgeable about these new tools and develop ways of integrating them into their curriculum"* (Koehler et al., 2011, p. 150). We may see evidence of this with *experienced* teachers in our MLM, whose classes showed higher CT compared to *new* teachers (Fig. 13). Such improved performance may be the byproduct of refined curriculum implementation, additional PD between year one and year two of the curriculum, and more careful consideration of how game design impacts CT. However, our quantitative focus limits our ability to move beyond conjectures to explain the mechanisms through which teacher experience may affect genre selection and GBL practices. In the future, more qualitative work can capture how teacher experience was refined through PD or how curriculum refinement may impact teachers' decision-making and classroom dynamics.

Despite this, we suggest that when introducing students to novel curriculum components, teachers take their time to engage with the new technology themselves while thoroughly reflecting on the implications of introducing such new components on content uptake and pedagogy (Tucker-Raymond et al., 2021). To that end, and based on the statistical evidence of teachers interacting with genres in student games, we recommend that future PD engaging CT and game design allow teachers to (1) understand how idiosyncratic design choices predict CT via game design (Troiano et al., 2020a) by exploring different genres of student-designed games related to their affordances in specific CT practices (2) engage with deep-play to engage with technology while reflecting on content and pedagogy (Koehler et al., 2011), having teachers spend at minimum 4–8 h designing games—and playtesting each other's games—and (3) become experienced with CT (Tagare, 2023) through game design and playtesting, ideally already in pre-service (İlic, 2021). This will allow them to understand how game design supports CT development, for which we recommend they reflect on how to balance their need to address CT learning objectives while supporting students' individual game design choices and preferences—without sapping *"all the spirit from the activity"* (Resnick & Rosenbaum, 2013).

Furthermore, as teachers engage in deep play, scaffolds should be created to support teachers' familiarity with advanced CT practices. For example, teachers need opportunities to explore logic, variables, and other advanced Scratch programming. A clear rubric should be created to highlight the importance of game genres and block usage in Scratch to encourage their understanding. Additionally, during PD, teachers should explore specific genres to become aware of how the type of game designed by their students can affect CT assessment. Consequently, to "level the computational playing field", teachers may ensure that students (1) acquire the foundations of programming and CT, including the use of logical statements, which may be scarce in certain game genres (e.g., *storytelling* and *quiz*), and (2) require that student games, regardless of genre, achieve at least a *developing* CT level in all CT dimensions.

4.4. Limitations

Our MLM was limited to a GBL, constructionist STEM curriculum. The secondary dataset employed in our MLM was pronounced overrepresented in specific demographics. For instance, White students were prevalent, creating imbalances in the dataset under analysis. Further, our gender analysis was also limited to our secondary dataset, which featured only males and females. Our results are also limited to Dr. Scratch. However, as shown in Fig. 4, the assessment of Dr. Scratch approximates a normal distribution in the data, which, although slightly skewed, suggests that the tool is measuring CT reliably. Prior work (Troiano et al., 2020a) showed that Dr. Scratch may inherently favor game genres (e.g., action) that require developing or proficient logical thinking (e.g., *if-else* blocks in Scratch), potentially assigning higher CT scores to students based on their design choices. As our MLM confirms these prior findings, we acknowledge that they might be contingent on the CT metrics we used to score CT in our analysis (i.e., Dr. Scratch) and highlight this as a limitation in our methodological approach. Future work may use other metric assessments (e.g., DWES; Chai et al. 2021) to confirm or challenge our results, as well as further scrutinize existing automated CT metrics for their inherent biases towards design choices in constructionist GBL. We used a random forest model as a data imputation strategy, limited to data availability, namely 55% of the dataset. Hence, it suffers from data imbalance, which we mitigated using stratified folds via cross-validation. Despite this, we could still classify correctly about 80% of the missing data, which produced a better performance than *softmax* (Wang et al., 2019), *decision tree* (Nikfalazar et al., 2020), and baseline imputation methods. While we used a proven gender classification, we acknowledge that these algorithms may be biased and that best practices for their ethical use are still debated (Lockhart et al., 2023).

Student pairs also limit our results, and we do not know to what extent each student has contributed to Scratch projects and their impact on CT scores. We showed how prior experience with Scratch and the game genre predicts CT. These results may be contingent on block usage in Scratch (Troiano et al., 2020a), which was not included in our MLM and should be inquired about in future work. While our findings highlight the predictive role of teacher experience and game genre in CT development, unexamined factors such as students' prior exposure to other forms of digital literacy or differences in instructional quality may also influence these outcomes, warranting further investigation. Last but not least, we did not inquire about remixes, *Bad Smells* (Vargas-Alba et al., 2019), or the use of "design shortcuts" like copy-pasting code snippets (Robles et al., 2017), all of which might have affected the MLM and resulting predictors of CT—we plan on extending our future work by including the aforementioned.

5. Conclusion

In this paper, we inquired about multiple factors that predict computational thinking (CT) performance in young students in a game-based learning (GBL) STEM curriculum. Specifically, we studied these factors in Scratch games designed by $n = 932$ students and assessed by Dr. Scratch. We performed a Multilevel Modeling (MLM) for our inquiry, using a *natural hierarchical structure* (i.e., the inherent organization of data into nested or hierarchical levels) and tested models consisting of four hierarchical levels for statistical significance in ascending order: (1) student-designed games, (2) cohort, (3) teachers, and (4) school. Through MLM, we revealed a hierarchical structure of predicting factors that included (1) teachers and their experience implementing the game-based STEM curriculum, (2) students' prior Scratch experience, (3) student-designed game genre, and (4) interaction between teacher experience and game genre. The influence of teachers emerged as prominent, who mediated the degree and the extent to which other factors, such as game genre and student experience with Scratch, predict CT score. With these findings, we advance ongoing efforts exploring factors predicting CT in young students and emphasize the hierarchical structure of the educational context in which they develop CT. Our discussion reveals implications for the design of CT and CS curricula, outlining avenues and needs to better support teachers in articulating CT uptake via GBL through further professional development.

CRediT authorship contribution statement

Giovanni M. Troiano: Writing – original draft, Validation, Supervision, Project administration, Methodology, Investigation, Funding acquisition, Data curation, Conceptualization. **Amir Abdollahi:** Writing – original draft, Visualization, Methodology, Formal analysis, Data curation. **Michael Cassidy:** Writing – review & editing, Supervision, Resources, Investigation, Funding acquisition, Data curation. **Gillian Puttick:** Writing – review & editing, Investigation, Funding acquisition, Data curation, Conceptualization. **Tiago Machado:** Writing – original draft, Validation, Formal analysis, Data curation. **Casper Harteveld:** Writing – original draft, Validation, Supervision, Project administration, Methodology, Investigation, Funding acquisition, Conceptualization.

Acknowledgments

This work was supported by the USA National Science Foundation (NSF) under the CSforALL federal research program (Award Number 2122707). We also want to thank Xiao Ma for helping with the data analysis.

Appendix A. Demographic survey

Pre-Survey 2017–2018 Student Demographics

We would like to find out a little bit more about you... Why are we asking you for information about yourself? This is a research project, funded by the National Science Foundation that is interested in learning. Knowing a bit more about you and the other students in the class will help us to improve the curriculum we are designing and testing in this project. If you wish not to take this survey, you do not have to. You can stop participation at any time.

The survey will take about 10 min to complete.

Student ID Number:

Teacher's Name:

1. What types of videogames do you like to play? *Please choose up to 3 types of videogames you like to play the most. If none, please click "Not applicable" and explain. If you play other types of videogames, please comment below.*

Preference	Role playing	Sports	Board/Card	Online multiplayer	Puzzles
Most	□	□	□	□	□
2nd Most	□	□	□	□	□
3rd Most	□	□	□	□	□

□ Not Applicable

2. On average, how many hours do you play video games each week? *Check one*

□ 0 h
□ 1–3 h
□ 4–6 h
□ 7–9 h
□ 10–12 h
□ 15–17 h
□ 18+ h

If other, please specify

3. What do you use computers for? *Please choose up to 3 uses for computers. If none, please click "Not applicable" and explain. If you use computers for another reason, please comment below.*

Preference	Finding information	Games	Watching videos	Socializing
Most	□	□	□	□
2nd Most	□	□	□	□
3rd Most	□	□	□	□

□ Not Applicable

If other, please specify

4. What are the kinds of technology you use? *Please choose up to 3 types of technology you use. If none, please click "Not applicable" and explain. If you use different types of technology, please comment below.*

□ Not Applicable

Preference	Phone	Tablet	Game consoles	Laptops	Desktops
Most	□	□	□	□	□
2nd Most	□	□	□	□	□
3rd Most	□	□	□	□	□

If other, please specify

5. Do you have Internet access at home? *Check one*

□ Yes
□ No

6. On average, how many hours do you use the Internet at home each week? *Check one*

□ 0 h
□ 1–3 h
□ 4–6 h
□ 7–9 h
□ 10–12 h
□ 15–17 h
□ 18+ hours

7. Have you made a computer game before? *Check one*

□ Yes
□ No

8. What is your gender? *If you do not wish to answer, you don't have to.*

□ Male
□ Female

9. Do you identify as Hispanic or Latino? *If you do not wish to answer, you don't have to.*

□ Yes
□ No

10. Race/ethnic group *If you do not wish to answer, you don't have to.*

□ American Indian or Alaskan Native
□ Asian
□ Black or African American
□ Native Hawaiian or Other Pacific Islander
□ White
□ Multiracial

If other, please specify

11. Is English your first language?

□ Yes
□ No

12. Do you speak a language other than English at home?

□ Yes
□ No

If yes, what language?

Appendix B. Prior experience with scratch

How much experience do you have using the software Scratch? *Check one*

□ Huh?
□ I have heard of it
□ Hour of code
□ Created a project
□ A lot of experience

Appendix C. MLM (without data imputation)

Table 7 presents a comparative analysis of three models in our MLM study, focusing on a refined dataset that excludes all the imputed data. Notably, *Model 1* demonstrates superior data fitting compared to the baseline *Model 0*, while *Model 2* does not exhibit an enhanced fit over *Model 1*. This finding contrasts with the results from the complete dataset, where *Model 2* outperformed *Model 1*. The limited dataset, excluding imputed data, results in a compromise of model robustness as it neglects approximately 51% of the data. The removal of projects with missing data on gender and experience does not hinder the model's prediction solely for these predictors but leads to a reduction in data points across other predictors of CT. Hence, especially for *Model 2* introducing the interaction of genre and teacher experience compared to *Model 1*, the limited dataset may lack sufficient data points for different genres and teacher experience levels, limiting the model's capacity to identify robust correlations and statistically significant trends in the interactive influence of these predictors on CT uptake. For instance, the absence of projects initiated under the guidance of new teachers in the storytelling genre in the new dataset precludes an examination of its interaction with teacher experience. Moreover, when employing AIC and BIC to assess the trade-off between the fitness of models and the model complexity, the performance in the new dataset diminishes more. AIC initially decreases from *Model 0* to *Model 1* but subsequently rises from *Model 1* to *Model 2* (for the complete version of the data discussed in the paper, it decreases entirely), and BIC exhibits an upward trend (at least initially decreasing for the complete version). These trends underscore the challenges of applying models to limited data when evaluating the trade-off between improved fit and increased complexity using AIC and BIC on the limited version of the data.

Additionally, in the new dataset, similar to the significant levels observed in the complete version, our thorough analysis of potential grouping variables revealed that only "teacher" reached statistical significance (Table 8) with a *p-value* of 4.3e−07. Subsequently, we examined the baseline model, denoted as *Model 0*, and two more complex models (*Model 1* and *Model 2*) within the new dataset. As discussed in the overall comparison of these models (Table 7) and detailed further (Tables 8 and 9), all models experienced a decrease in fit when applied to the new limited dataset. Certain predictors, such as student experience and specific genres, maintained their significant predictive power for CT uptake even in this restricted data, underscoring their robust correlation (though compromised compared to the complete dataset). Meanwhile, some less robust predictors became statistically insignificant in this updated version.

In summary, a comprehensive assessment of model fitness, considering both AIC and BIC, underscores the significance of utilizing the more complete version of the data, even when encountering missing data for gender and experience levels of students in certain projects. These models, particularly *Model 2* with its consideration of interaction effects, provide more valuable insights into the factors influencing CT when applied to the complete data. Simultaneously, they address concerns about overfitting, as evidenced by AIC and BIC assessments. This comparative analysis reinforces the rationale for favoring the complete dataset, as restricting the analysis to projects with complete data not only hinders model performance but also discards valuable information available for those projects.

Data availability

Data will be made available on request.

Table 7
The three MLM models and their statistical comparisons — without imputed data.

Model	npar	AIC	BIC	logLik	Chisq	Df	Pr(>Chisq)
Model 0	3	1338.8	1349.8	−666.40			
Model 1	19	1309.8	1379.5	−635.89	61.017	16	**3.52e−07** ***
Model 2	28	1320.5	1423.3	−632.27	7.237	9	0.613

Table 8
P-value and ICC for potential attributes in grouping observations (levels)- without imputed data.

Independent variable	P value	ICC
School	0.92	0.05
Teacher	**4.3e−07** ***	**0.10**
Cohort	1	0.01

*** Significant at $p < 0.001$.

Table 9
The resulting MLM models, fixed effects (independent variables), and statistical values — without imputed data.

Fixed effects	Model 0: *Baseline*		Model 1: *Intermediate*		Model 2: *Final*	
	Estimate	P-value	Estimate	P-value	Estimate	P-value
Intercept	13.96	**4.06e−13** ***	18.74	**2.81e−05** ***	16.36	**6.07e−05** ***
Main Effects						
Experience-2			0.29	0.409	0.29	0.419
Experience-3			0.56	0.165	0.55	0.174
Experience-4			1.24	**0.006** **	1.11	**0.015** *
Experience-5			1.71	**0.003** **	1.57	**0.006** **
Teacher Experience-New			−1.07	0.127	1.97	0.243
Genre-Clicker			−0.41	0.452	1.92	0.232
Genre-Maze			−0.02	0.975	2.86	0.097
Genre-Platform			−0.54	0.462	2.48	0.129
Genre-Pong			−0.97	0.084	1.93	0.228
Genre-Puzzle			0.49	0.564	3.58	**0.043** *
Genre-Quiz			−2.38	**9.35e−05** ***	0.16	0.923
Genre-Shooter			1.96	0.071	3.68	0.085
Genre-Simulation			0.37	0.606	3.35	**0.040** *
Genre-Storytelling			−1.17	0.112	−1.59	**0.034** *
Genre-Swipe elimination			−1.50	**0.048** *	1.35	0.492
Cohort-2			−1.35	0.063	−1.45	0.070
Interaction Effects						
Teacher Experience-New: Genre-Clicker					−2.62	0.127
Teacher Experience-New: Genre-Maze					−3.39	0.073
Teacher Experience-New: Genre-Platform					−4.14	**0.034** *
Teacher Experience-New: Genre-Pong					−3.43	**0.048** *
Teacher Experience-New: Genre-Puzzle					−4.01	0.061
Teacher Experience-New: Genre-Quiz					−2.83	0.111
Teacher Experience-New: Genre-Shooter					−1.28	0.606
Teacher Experience-New: Genre-Simulation					−4.19	**0.037** *
Teacher Experience-New: Genre-Storytelling						
Teacher Experience-New: Genre-Swipe elimination					−3.32	0.121
	Model 0		Model 1		Model 2	
Random Effects	**logLik**	**P-value**	**logLik**	**P-value**	**logLik**	**P-value**
Level 2: Teacher	−679.2	**4.31e−07** ***	−639.3	**0.008** **	−636.2	**0.005** **

References

Angeli, C., & Valanides, N. (2020). Developing young children’s computational thinking with educational robotics: An interaction effect between gender and scaffolding strategy. *Computers in Human Behavior*, *105*, Article 105954.

Ardito, G., Czerkawski, B., & Scollins, L. (2020). Learning computational thinking together: Effects of gender differences in collaborative middle school robotics program. *TechTrends*, *64*(3), 373–387. http://dx.doi.org/10.1007/s11528-019-00461-8, URL http://link.springer.com/10.1007/s11528-019-00461-8.

Atman Uslu, N. (2023). How do computational thinking self-efficacy and performance differ according to secondary school students’ profiles? The role of computational identity, academic resilience, and gender. *Education and Information Technologies*, *28*(5), 6115–6139. http://dx.doi.org/10.1007/s10639-022-11425-6, URL https://link.springer.com/10.1007/s10639-022-11425-6.

Barr, V., & Stephenson, C. (2011). Bringing computational thinking to K-12: What is involved and what is the role of the computer science education community? *ACM Inroads*, *2*(1), 48–54.

Baser, M. (2013). *Attitude, gender and achievement in computer programming*: *Technical report*, (pp. 248–255). ISSN: 1990-9233 Issue: 2 Publication Title: Online Submission Volume: 14 ERIC Number: ED542330, URL https://eric.ed.gov/?id=ED542330.

Beyer, S. (2014). Why are women underrepresented in computer science? Gender differences in stereotypes, self-efficacy, values, and interests and predictors of future CS course-taking and grades. *Computer Science Education*, *24*(2–3), 153–192.

Blevins, C., & Mullen, L. (2015). Jane, John... Leslie? A historical method for algorithmic gender prediction. *DHQ: Digital Humanities Quarterly*, *9*(3).

Camacho, J., & Ferrer, A. (2012). Cross-validation in PCA models with the element-wise k-fold (ekf) algorithm: theoretical aspects. *Journal of Chemometrics*, *26*(7), 361–373.

Cassidy, M., Tucker-Raymond, E., & Puttick, G. (2020). Distributing expertise to integrate computational thinking practices. *Science Scope*, *43*(7), 18–21.

Chai, X., Sun, Y., Luo, H., & Guizani, M. (2021). DWES: a dynamic weighted evaluation system for scratch based on computational thinking. *IEEE Transactions on Emerging Topics in Computing*, *10*(2), 917–932.

Cheema, J. R. (2014). A review of missing data handling methods in education research. *Review of Educational Research*, *84*(4), 487–508.

Chongo, S., Osman, K., & Nayan, N. A. (2020). Level of computational thinkinglevel of computational thinking skills among secondary science student: Variation across gender and mathematics achievement skills among secondary science student: Variation across gender and mathematics achievement. *Science Education International*, *31*(2), 159–163. http://dx.doi.org/10.33828/sei.v31.i2.4, URL http://www.icaseonline.net/journal/index.php/sei/article/view/204.

Czerkawski, B. C., & Lyman, E. W. (2015). Exploring issues about computational thinking in higher education. *TechTrends*, *59*(2), 57–65. http://dx.doi.org/10.1007/s11528-015-0840-3, URL http://link.springer.com/10.1007/s11528-015-0840-3.

Dedrick, R. F., Ferron, J. M., Hess, M. R., Hogarty, K. Y., Kromrey, J. D., Lang, T. R., Niles, J. D., & Lee, R. S. (2009). Multilevel modeling: A review of methodological issues and applications. *Review of Educational Research*, *79*(1), 69–102.

Del Olmo-Muñoz, J., Cózar-Gutiérrez, R., & González-Calero, J. A. (2020). Computational thinking through unplugged activities in early years of primary education. *Computers & Education*, *150*, Article 103832. http://dx.doi.org/10.1016/j.compedu.2020.103832, URL https://linkinghub.elsevier.com/retrieve/pii/S0360131520300348.

Delal, H., & Oner, D. (2020). Developing middle school students' computational thinking skills using unplugged computing activities. *Informatics in Education*, *19*(1), 1–13. http://dx.doi.org/10.15388/infedu.2020.01, URL https://infedu.vu.lt/doi/10.15388/infedu.2020.01.

Durak, H. Y., & Saritepeci, M. (2018). Analysis of the relation between computational thinking skills and various variables with the structural equation model. *Computers & Education*, *116*, 191–202. http://dx.doi.org/10.1016/j.compedu.2017.09.004, URL https://linkinghub.elsevier.com/retrieve/pii/S0360131517302087.

Freeman, J., Magerko, B., Edwards, D., Mcklin, T., Lee, T., & Moore, R. (2019). EarSketch: engaging broad populations in computing through music. *Communications of the ACM*, *62*(9), 78–85.

Funke, A., & Geldreich, K. (2017). Gender differences in scratch programs of primary school children. In *Proceedings of the 12th workshop on primary and secondary computing education* (pp. 57–64). Nijmegen Netherlands: ACM, http://dx.doi.org/10.1145/3137065.3137067, URL https://dl.acm.org/doi/10.1145/3137065.3137067.

Gao, H., Hasenbein, L., Bozkir, E., Göllner, R., & Kasneci, E. (2022). Exploring gender differences in computational thinking learning in a VR classroom: Developing machine learning models using eye-tracking data and explaining the models. *International Journal of Artificial Intelligence in Education*, http://dx.doi.org/10.1007/s40593-022-00316-z, URL https://link.springer.com/10.1007/s40593-022-00316-z.

Gelman, A. (2006). Multilevel (hierarchical) modeling: what it can and cannot do. *Technometrics*, *48*(3), 432–435.

Greig, A. D., Taylor, M. J., & MacKay, T. (2007). *Doing research with children*. Sage.

Guggemos, J. (2021). On the predictors of computational thinking and its growth at the high-school level. *Computers & Education*, *161*, Article 104060. http://dx.doi.org/10.1016/j.compedu.2020.104060, URL https://linkinghub.elsevier.com/retrieve/pii/S036013152030258X.

Harteveld, C. (2011). *Triadic game design: Balancing reality, meaning and play*. Springer Science & Business Media.

Heintz, S., & Law, E. L.-C. (2015). The game genre map: A revised game classification. In *CHI PLAY '15*, *Proceedings of the 2015 annual symposium on computer-human interaction in play* (pp. 175–184). New York, NY, USA: ACM, http://dx.doi.org/10.1145/2793107.2793123, URL http://doi.acm.org/10.1145/2793107.2793123.

Honey, M. (2013). *Design, make, play: Growing the next generation of STEM innovators*. Routledge.

Hox, J., Moerbeek, M., & Van de Schoot, R. (2017). *Multilevel analysis: Techniques and applications*. Routledge.

Hox, J., & Roberts, J. K. (2011). *Handbook of advanced multilevel analysis*. Psychology Press.

Hutchins, N. M., Zhang, N., & Biswas, G. (2017). The role gender differences in computational thinking confidence levels plays in STEM applications. In *The international conference on computational thinking education (CTE'17). the education university of Hong Kong, Hong Kong* (pp. 34–38).

İlic, U. (2021). The impact of scratch-assisted instruction on computational thinking (CT) skills of pre-service teachers. *International Journal of Research in Education and Science (IJRES)*.

Israel-Fishelson, R., Hershkovitz, A., Eguíluz, A., Garaizar, P., & Guenaga, M. (2021). The associations between computational thinking and creativity: The role of personal characteristics. *Journal of Educational Computing Research*, *58*(8), 1415–1447. http://dx.doi.org/10.1177/0735633120940954, URL http://journals.sagepub.com/doi/10.1177/0735633120940954.

James, G., Witten, D., Hastie, T., Tibshirani, R., et al. (2013). *An introduction to statistical learning*: *vol. 112*, Springer.

Jin, Y., Sun, J., Ma, H., & Wang, X. (2021). The impact of different types of scaffolding in project-based learning on girls' computational thinking skills and self-efficacy. In *2021 tenth international conference of educational innovation through technology* (pp. 362–366). Chongqing, China: IEEE, http://dx.doi.org/10.1109/EITT53287.2021.00077, URL https://ieeexplore.ieee.org/document/9694003/.

Kafai, Y. (1995). *Minds in play: Computer game design as a context for children's learning*. Routledge, Google-Books-ID: Ocyllxa8ZjkC.

Kafai, Y. B., & Burke, Q. (2017). Computational participation: Teaching kids to create and connect through code. *Emerging Research, Practice, and Policy on Computational Thinking*, 393–405.

Karimi, F., Wagner, C., Lemmerich, F., Jadidi, M., & Strohmaier, M. (2016). Inferring gender from names on the web: A comparative evaluation of gender detection methods. In *Proceedings of the 25th international conference companion on world wide web* (pp. 53–54).

Khine, M. S. (2022). *Methodology for multilevel modeling in educational research: Concepts and applications*. Springer Nature.

Kiss, G. (2010). A comparison of programming skills by genders of hungarian grammar school students. In *2010 7th international conference on ubiquitous intelligence & computing and 7th international conference on autonomic & trusted computing* (pp. 24–30). Xi'an, Shaanxi, China: IEEE, http://dx.doi.org/10.1109/UIC-ATC.2010.83, URL http://ieeexplore.ieee.org/document/5667105/.

Koehler, M. J., Mishra, P., Bouck, E. C., DeSchryver, M., Kereluik, K., Shin, T. S., & Wolf, L. G. (2011). Deep-play: Developing TPACK for 21st century teachers. *International Journal of Learning Technology*, *6*(2), 146–163.

Kong, S.-C., Chiu, M. M., & Lai, M. (2018). A study of primary school students' interest, collaboration attitude, and programming empowerment in computational thinking education. *Computers & Education*, *127*, 178–189. http://dx.doi.org/10.1016/j.compedu.2018.08.026, URL https://linkinghub.elsevier.com/retrieve/pii/S0360131518302367.

Kong, S., & Lai, M. (2022). Computational identity and programming empowerment of students in computational thinking development. *British Journal of Educational Technology*, *53*(3), 668–686. http://dx.doi.org/10.1111/bjet.13175, URL https://onlinelibrary.wiley.com/doi/10.1111/bjet.13175.

Korucu, A. T., Gencturk, A. T., & Gundogdu, M. M. (2017). Examination of the Computational Thinking Skills of Students.

Latifah, S., Diani, R., & Malik, S. L. M. (2022). ICARE model (introduction, connection, application, reflection, extension) in physics learning: Analysis of its effect on students' computational thinking skills based on gender. *Jurnal Penelitian & Pengembangan Pendidikan Fisika*, *8*(2), 229–240. http://dx.doi.org/10.21009/1.08205, URL http://journal.unj.ac.id/unj/index.php/jpppf/article/view/28035.

Lei, H., Chiu, M. M., Li, F., Wang, X., & Geng, Y.-j. (2020). Computational thinking and academic achievement: A meta-analysis among students. *Children and Youth Services Review*, *118*, Article 105439. http://dx.doi.org/10.1016/j.childyouth.2020.105439, URL https://linkinghub.elsevier.com/retrieve/pii/S0190740920311725.

Leyland, A. H., & Groenewegen, P. P. (2020). *Multilevel modelling for public health and health services research: health in context*. Springer Nature.

Lockhart, J. W., King, M. M., & Munsch, C. (2023). Name-based demographic inference and the unequal distribution of misrecognition. *Nature Human Behaviour*, 1–12.

Lockwood, T. (2010). *Design thinking: Integrating innovation, customer experience, and brand value*. Simon and Schuster.

Markus, K. A. (2012). *Principles and practice of structural equation modeling by Rex B. Kline*. Taylor & Francis.

Michell, D., Szabo, C., Falkner, K., & Szorenyi, A. (2018). Towards a socio-ecological framework to address gender inequity in computer science. *Computers & Education*, *126*, 324–333.

Mihaljević, H., Tullney, M., Santamaría, L., & Steinfeldt, C. (2019). Reflections on gender analyses of bibliographic corpora. *Frontiers in Big Data*, *2*, 29.

Moon, H., & Cheon, J. T. (2023). An investigation of affective factors influencing computational thinking and problem-solving.

Moreno-León, J., Robles, G., & Román-González, M. (2015). Dr. Scratch: Automatic analysis of scratch projects to assess and foster computational thinking. *RED. Revista de Educación A Distancia*, *15*(46).

Moreno-León, J., Robles, G., & Román-González, M. (2017). Towards data-driven learning paths to develop computational thinking with scratch. *IEEE Transactions on Emerging Topics in Computing*, *8*(1), 193–205.

Nikfalazar, S., Yeh, C.-H., Bedingfield, S., & Khorshidi, H. A. (2020). Missing data imputation using decision trees and fuzzy clustering with iterative learning. *Knowledge and Information Systems*, *62*, 2419–2437.

Niousha, R., Saito, D., Washizaki, H., & Fukazawa, Y. (2022). Scratch project analysis: Relationship between gender and computational thinking skill. In *2022 IEEE international conference on teaching, assessment and learning for engineering* (pp. 567–571). Hung Hom, Hong Kong: IEEE, http://dx.doi.org/10.1109/TALE54877.2022.00099, URL https://ieeexplore.ieee.org/document/10148435/.

Papert, S. (1980). *Mindstorms: Children, computers, and powerful ideas*. New York, NY, USA: Basic Books, Inc..

Park, Y., & Shin, Y. (2019). Comparing the effectiveness of scratch and app inventor with regard to learning computational thinking concepts. *Electronics*, *8*(11), 1269.

Peckham, J., Harlow, L. L., Stuart, D. A., Silver, B., Mederer, H., & Stephenson, P. D. (2007). Broadening participation in computing: issues and challenges. *ACM SIGCSE Bulletin*, *39*(3), 9–13.

Peugh, J. L. (2010). A practical guide to multilevel modeling. *Journal of School Psychology*, *48*(1), 85–112.

Piaget, J. (1959). The language and thought of the child (1926). New York and Scarborough: Meridien for NAL.

Polat, E., Hopcan, S., Kucuk, S., & Sisman, B. (2021). A comprehensive assessment of secondary school students' computational thinking skills. *British Journal of Educational Technology*, *52*(5), 1965–1980. http://dx.doi.org/10.1111/bjet.13092, URL https://onlinelibrary.wiley.com/doi/10.1111/bjet.13092.

Prana, G. A. A., Ford, D., Rastogi, A., Lo, D., Purandare, R., & Nagappan, N. (2021). Including everyone, everywhere: Understanding opportunities and challenges of geographic gender-inclusion in oss. *IEEE Transactions on Software Engineering*, *48*(9), 3394–3409.

Puttick, G., & Tucker-Raymond, E. (2018). Building systems from scratch: An exploratory study of students learning about climate change. *Journal of Science Education and Technology*, *27*, 306–321.

Relkin, E., De Ruiter, L., & Bers, M. U. (2020). TechCheck: Development and validation of an unplugged assessment of computational thinking in early childhood education. *Journal of Science Education and Technology*, *29*(4), 482–498. http://dx.doi.org/10.1007/s10956-020-09831-x, URL https://link.springer.com/10.1007/s10956-020-09831-x.

Relkin, E., de Ruiter, L. E., & Bers, M. U. (2021). Learning to code and the acquisition of computational thinking by young children. *Computers & Education*, *169*, Article 104222.

Resnick, M., Maloney, J., Monroy-Hernández, A., Rusk, N., Eastmond, E., Brennan, K., Millner, A., Rosenbaum, E., Silver, J., Silverman, B., et al. (2009). Scratch: programming for all. *Communications of the ACM*, *52*(11), 60–67.

Resnick, M., & Rosenbaum, E. (2013). Designing for tinkerability. In *Design, make, play* (pp. 163–181). Routledge.

Rijke, W. J., Bollen, L., Eysink, T. H. S., & Tolboom, J. L. J. (2018). Computational thinking in primary school: An examination of abstraction and decomposition in different age groups. *Informatics in Education*, *17*(1), 77–92. http://dx.doi.org/10.15388/infedu.2018.05, URL https://infedu.vu.lt/doi/10.15388/infedu.2018.05.

Roberts, J. K. (2004). An introductory primer on multilevel and hierarchical linear modeling.. *Learning Disabilities: A Contemporary Journal*, *2*(1), 30–38.

Robles, G., Moreno-León, J., Aivaloglou, E., & Hermans, F. (2017). Software clones in scratch projects: On the presence of copy-and-paste in computational thinking learning. In *2017 IEEE 11th international workshop on software clones* (pp. 1–7). IEEE.

Rojas López, A., & García-Peñalvo, F. J. (2021). Initial performance analysis in the evaluation of computational thinking from a gender perspective in higher education. In *Ninth international conference on technological ecosystems for enhancing multiculturality* (pp. 109–114). Barcelona Spain: ACM, http://dx.doi.org/10.1145/3486011.3486429, URL https://dl.acm.org/doi/10.1145/3486011.3486429.

Román-González, M., Pérez-González, J.-C., & Jiménez-Fernández, C. (2017). Which cognitive abilities underlie computational thinking? Criterion validity of the computational thinking test. *Computers in Human Behavior*, *72*, 678–691. http://dx.doi.org/10.1016/j.chb.2016.08.047, URL https://linkinghub.elsevier.com/retrieve/pii/S0747563216306185.

Rosebery, A. S., & Puttick, G. M. (1997). Teacher professional development as situated inquiry: A case study in science education. Center for the development of teaching paper series.

Santamaría, L., & Mihaljević, H. (2018). Comparison and benchmark of name-to-gender inference services. *PeerJ Computer Science*, *4*, Article e156.

Sax, L. J., Lehman, K. J., Jacobs, J. A., Kanny, M. A., Lim, G., Monje-Paulson, L., & Zimmerman, H. B. (2017). Anatomy of an enduring gender gap: The evolution of women's participation in computer science. *The Journal of Higher Education*, *88*(2), 258–293. http://dx.doi.org/10.1080/00221546.2016.1257306, Publisher: Routledge.

Scopatz, A., & Huff, K. D. (2015). *Effective computation in physics: Field guide to research with Python*. " O'Reilly Media, Inc.".

Sebo, P. (2021). Performance of gender detection tools: a comparative study of name-to-gender inference services. *Journal of the Medical Library Association: JMLA*, *109*(3), 414.

Seiter, L., & Foreman, B. (2013). Modeling the learning progressions of computational thinking of primary grade students. In *Proceedings of the ninth annual international ACM conference on international computing education research* (pp. 59–66). San Diego San California USA: ACM, http://dx.doi.org/10.1145/2493394.2493403, URL https://dl.acm.org/doi/10.1145/2493394.2493403.

Serkan, M., & Karalar, H. (2018). Gender differences in middle school students' attitudes and self-efficacy perceptions towards mblock programming. *European Journal of Educational Research*, *7*(4), 925–933. http://dx.doi.org/10.12973/eu-jer.7.4.925, URL https://eu-jer.com/gender-differences-in-middle-school-students-attitudes-and-self-efficacy-perceptions-towards-mblock-programming.

Song, Q., & Shepperd, M. (2007). Missing data imputation techniques. *International Journal of Business Intelligence and Data Mining*, *2*(3), 261–291.

Sovey, S., Osman, K., & Matore, M. E. E. M. (2022). Gender differential item functioning analysis in measuring computational thinking disposition among secondary school students. *Frontiers in Psychiatry*, *13*, Article 1022304. http://dx.doi.org/10.3389/fpsyt.2022.1022304, URL https://www.frontiersin.org/articles/10.3389/fpsyt.2022.1022304/full.

Stein, C. (2004). Botball robotics and gender differences in middle school teams. In *2004 annual conference proceedings* (pp. 9.262.1–9.262.10). Salt Lake City, Utah: ASEE Conferences, http://dx.doi.org/10.18260/1-2–13534, URL http://peer.asee.org/13534.

Stoilescu, D., & Egodawatte, G. (2010). Gender differences in the use of computers, programming, and peer interactions in computer science classrooms. *Computer Science Education*, *20*(4), 283–300. http://dx.doi.org/10.1080/08993408.2010.527691, URL http://www.tandfonline.com/doi/abs/10.1080/08993408.2010.527691.

Sun, L., Hu, L., & Zhou, D. (2022). The bidirectional predictions between primary school students' STEM and language academic achievements and computational thinking: The moderating role of gender. *Thinking Skills and Creativity*, *44*, Article 101043. http://dx.doi.org/10.1016/j.tsc.2022.101043, URL https://linkinghub.elsevier.com/retrieve/pii/S1871187122000463.

Tagare, D. (2023). Factors that predict K-12 teachers' ability to apply computational thinking skills. *ACM Transactions on Computing Education*.

Tang, F., & Ishwaran, H. (2017). Random forest missing data algorithms. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, *10*(6), 363–377.

Troiano, G. M., Chen, Q., Alba, Á. V., Robles, G., Smith, G., Cassidy, M., Tucker-Raymond, E., Puttick, G., & Harteveld, C. (2020). Exploring how game genre in student-designed games influences computational thinking development. In *Proceedings of the 2020 CHI conference on human factors in computing systems* (pp. 1–17).

Troiano, G. M., Schouten, D., Cassidy, M., Tucker-Raymond, E., Puttick, G., & Harteveld, C. (2020). Ice paddles, CO2 invaders, and exploding planets: how young students transform climate science into serious games. In *Proceedings of the annual symposium on computer-human interaction in play* (pp. 534–548).

Troiano, G. M., Snodgrass, S., Argımak, E., Robles, G., Smith, G., Cassidy, M., Tucker-Raymond, E., Puttick, G., & Harteveld, C. (2019). Is my game OK dr. Scratch? Exploring programming and computational thinking development via metrics in student-designed serious games for STEM. In *Proceedings of the 18th ACM international conference on interaction design and children* (pp. 208–219).

Tucker-Raymond, E., Cassidy, M., & Puttick, G. (2021). Science teachers can teach computational thinking through distributed expertise. *Computers & Education*, *173*, Article 104284.

Tucker-Raymond, E., Puttick, G., Cassidy, M., Harteveld, C., & Troiano, G. M. (2019). "I Broke Your Game!": critique among middle schoolers designing computer games about climate change. *International Journal of STEM Education*, *6*, 1–16.

Vargas-Alba, Á., Troiano, G. M., Chen, Q., Harteveld, C., & Robles, G. (2019). Bad smells in scratch projects: A preliminary analysis. In *TACKLE@ EC-TEL*.

Vieira, J. M. F. (2020). Learning trajectories visualizations of visual programming on the computational thinking context.

Wang, J., Hong, H., Ravitz, J., & Ivory, M. (2015). Gender differences in factors influencing pursuit of computer science and related fields. In *Proceedings of the 2015 ACM conference on innovation and technology in computer science education* (pp. 117–122).

Wang, S., Li, B., Yang, M., & Yan, Z. (2019). Missing data imputation for machine learning. In *IoT as a service: 4th EAI international conference, ioTaaS 2018, xi'an, China, November 17–18, 2018, proceedings 4* (pp. 67–72). Springer.

Weintrop, D., Beheshti, E., Horn, M., Orton, K., Jona, K., Trouille, L., & Wilensky, U. (2016). Defining computational thinking for mathematics and science classrooms. *Journal of Science Education and Technology*, *25*, 127–147.

Werner, L., Denner, J., Bliesner, M., & Rex, P. (2009). Can middle-schoolers use storytelling alice to make games? Results of a pilot study. In *Proceedings of the 4th international conference on foundations of digital games* (pp. 207–214).

Werner, L., Denner, J., Campe, S., & Kawamoto, D. C. (2012). The fairy performance assessment: measuring computational thinking in middle school. In *Proceedings of the 43rd ACM technical symposium on computer science education* (pp. 215–220). Raleigh North Carolina USA: ACM, http://dx.doi.org/10.1145/2157136.2157200, URL https://dl.acm.org/doi/10.1145/2157136.2157200.

Werner, L. L., Hanks, B., & McDowell, C. (2004). Pair-programming helps female computer science students. *Journal on Educational Resources in Computing*, *4*(1), 4. http://dx.doi.org/10.1145/1060071.1060075, URL https://dl.acm.org/doi/10.1145/1060071.1060075.