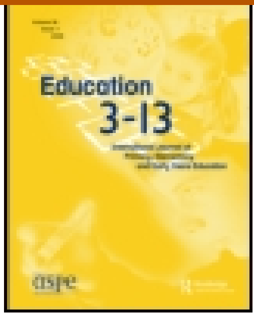
Routledge
Taylor & Francis Group



Education 3-13

International Journal of Primary, Elementary and Early Years Education

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/rett20

Exploring gender differences in primary school computer programming classes: a study in an English state-funded urban school

Colin B. Price & Ruth Price-Mohr

To cite this article: Colin B. Price & Ruth Price-Mohr (2021): Exploring gender differences in primary school computer programming classes: a study in an English state-funded urban school, Education 3-13, DOI: 10.1080/03004279.2021.1971274

To link to this article: https://doi.org/10.1080/03004279.2021.1971274

EDUCATION 3-13
https://doi.org/10.1080/03004279.2021.1971274

OPEN ACCESS Check for updates

Exploring gender differences in primary school computer programming classes: a study in an English state-funded urban school

Colin B. Price and Ruth Price-Mohr

Department of Computing, University of Worcester, Worcester, UK

ABSTRACT

This study investigates computer programming ('coding') activities of Primary School Children; we ask if there is evidence of gender differences in their coding activities. The research took place in an English urban school with around 180 children on role, mostly from a middle-class social background. The study involved a class of 17 boys and 15 girls aged 10–11 years. Teaching was delivered by the class teacher using our 'WeeBee engine', where children code animated stories using the professional text-based language Java. We first review relevant literature about gender differences to develop criteria for our analysis. We assess the children's code, their process of coding and the quality of their final animated stories. Our findings strongly suggest there are no gender differences in coding ability and in the quality of stories created. We suggest that practitioners should not assume that gender differences exist in this context, and they should not adapt their teaching to gender. The WeeBee engine is established as being gender neutral and we recommend its use by practitioners.

ARTICLE HISTORY

Received 30 September 2020
Accepted 26 July 2021

KEYWORDS

Primary school; elementary school; computer programming; gender differences

Introduction

Over the past few years, the subject of Computer Science (CS) has entered compulsory primary school education. This started in 2013 in England and has now spread globally including Finland, Israel, Estonia, and Japan. The thinking behind this is multi-fold, first as a valid educational activity in its own right; the educational value of CS, and especially coding (programming), is to enable children to become producers, rather than consumers, of digital artefacts that pervade their lives (Husbye et al. 2012; Wohlwend 2015), indeed coding can be viewed as a new form of literacy (Burnett 2016). Other benefits of coding in primary schools have been noted, such as improvement in planning and inhibition tasks (Arfé et al. 2019) and improved levels of mathematical thinking (Miller 2019). Our experience from working with hundreds of children over the past years has taught us that most children find coding a motivating, fun, and rewarding experience. Our research focusses on the use of Java code to compose animated stories that can be taken forward into Literacy classes.

The statutory inclusion of CS in primary and secondary education requires us to address the question of the under-representation of females in higher education and the CS workforce. While space does not allow a discussion of the under-representation issue here, we direct the interested reader to Wang and Degol (2017) who provide both a review and potential explanations for this situation.

CONTACT Colin B. Price c.price@worc.ac.uk
Supplemental data for this article can be accessed at https://doi.org/10.1080/03004279.2021.1971274

We feel it is important to investigate potential gender issues at all educational levels, especially when a new subject, such as CS, is mandated. In this paper, we report on coding activities undertaken and computer programs created by a class of 17 girls and 15 boys in a middle-class urban primary school. This paper is structured as follows. In the remainder of this section, we call upon the literature concerning gender issues in coding and in reading and writing, and also provide a short overview of our approach to coding. We then report our research methodology that consists of three measures: choice of code methods used; the *process* of coding; and the quality of the animation produced. Our findings section draws on the literature discussed; this is followed by a discussion and conclusion.

Background – gender and coding

There have been several coding platforms developed for young children aimed to provide a motivating and fun learning experience; these include Scratch, Alice, and Kodu. Some approaches have been developed with gender explicitly in mind, focussing on girls, such as Storytelling Alice (Kelleher, Pausch, and Kiesler 2007), and the Arduino Lily-pad where children can produce and code interactive wearable textiles (Buechley et al. 2008). These and other studies have focussed on approaches for decreasing the reported gender gap in participation in CS, such as developing positive attitudes and confidence towards coding (Kalelioglu 2015; Çakir et al. 2017). Scratch (Resnick et al. 2009) is promoted as the most suitable platform for young children. Here they assemble code by dragging and dropping functional blocks, rather than typing text (CAS 2015/2017). Our engine differs substantially in that children code by writing text, using the professional language Java; our prior research has confirmed that children are quite capable of doing this (Price and Price-Mohr 2018a, 2018b).

There have been only a few studies which investigate gender issues in computer science; this is perhaps expected as the research field is relatively new. An investigation into the use of Scratch by 4th-grade children (ages 9–10) showed that boys ($n = 32$) and girls ($n = 26$) made different use of blocks, with boys preferring ‘motion’ blocks, and girls preferring ‘look’ blocks. Looks blocks relate to speech or changing the character’s appearance, while motion blocks are concerned with movement (Funke and Geldreich 2017). The authors also found that their final artefacts differed, with most girls producing stories and most boys producing games.

A crucial pedagogical element is how children are organised in the classroom. This includes ‘pair programming’ where children working in pairs take turns in typing in code and critiquing the resulting animation. This has been shown to have benefit in engaging children from under-represented groups (Denner and Campe 2018). The question arises, how to mix the genders in the pairs. One study investigated the combinations of GG, BG and BB, and found the most effective pairing was GG with the BB least effective (Tsan, Boyer, and Lynch 2016). This suggests that practitioners need to reflect on pair grouping for the particular children in their class.

There is also the suggestion that girls approach a programming task by first thinking and planning while boys launch into the coding process immediately. Papavlasopoulou, Sharma, and Giannakos (2020) report a study of 69 boys and 36 girls in Norwegian after-school clubs using Scratch to produce a game; observations of the children showed that girls started the activity by thinking and planning the game, whereas boys immediately started experimenting with various block actions without having a concrete plan. Another study presents the teacher’s perspective where 63 Bavarian CS teachers were invited to report on their experiences of their children in programming (Funke et al. 2015). Here 44% of the teachers surveyed reported that ‘Boys just start working, girls think about it first’, 29% reported ‘Boys want to try more, test, find things out’ while 19% were not aware of any gender differences.

Recent research has attempted to triangulate eye-tracking behaviour and interview results (Papavlasopoulou, Sharma, and Giannakos 2020). Eye-tracking was used to identify the various elements of the Scratch screen, such as block commands and the assembled block program. The authors argue that the combination of quantitative eye-tracking behaviour with qualitative interview data will establish an objective measure for gender difference. They report that there is no difference in their behaviour based on their gaze and they conclude there is no difference in the children's cognitive processes. We feel the link between gaze and cognitive processes seems hard to establish, and present below what we think is a much more robust instrument, directly monitoring the *process* of children's coding as they type their code.

The WeeBee engine is rather unique in that it involves text, and therefore explicit reading and writing (typing) of lines of code with a particular grammar. This suggests we should briefly consider gender differences in reading and writing as reported in the literature.

Background – gender and literacy

Literacy research often quotes the effect size (0.0–1.0), which quantifies the size of an intervention's result. The most common measure of effect size is Cohen's *d*, see Hyde (2014) for a discussion. Values of $d < 0.01$ are trivial, 0.20 small, 0.50 moderate and 0.80 large. Here we assume that a positive *d* indicates there is a gender difference in the favour of boys and a negative *d* in favour of girls.

Within Literacy, there are two well-researched hypotheses, first is the gender similarities hypothesis (Hyde 2005) which claims there is little difference between boys and girls on most psychological variables. Hyde presents strong evidence for this hypothesis (Hyde 2005, 2014) but does note that there are exceptions. Exceptions relevant to this paper are reading achievement $d = -0.41$ to -0.46 (Reilly 2015), verbal fluency (Hyde and Linn 1988), writing productivity $d = -0.25$ and writing quality $d = -0.26$ (Adams and Simmons 2019). All these effect sizes indicate that girls perform better on these measures.

Some behavioural measures have been reported as exceptions to the rule such as effortful control $d = -0.41$ (Else-Quest, Hyde, and Linn 2010), self-regulation $d = -0.54$ (Walker and Berthelsen 2017), agreeableness $d = -1.07$ (Feingold 1994), 'things versus people' $d = 0.92$ (Su, Rounds, and Armstrong 2009), sensation seeking $d = 0.41$ (Cross, Copping, and Campbell 2011) work attitude $d = -0.44$ and competition $d = 0.39$ (Driessen and van Langen 2013). These results suggest girls have a more self-disciplined approach to learning while boys tend to be more competitive and sensation seeking.

The second hypothesis is that of greater male variability, where scores on measures show a larger range of values for boys than girls (Feingold 1992; Machin and Pekkarinen 2008). Interestingly, a comparison of STEM and non-STEM subjects confirms this hypothesis but also suggests that both means, and variance of grades, are smaller in STEM than non-STEM subjects (O'Dea et al. 2018). Given the greater male variability, it is important that in looking for potential gender differences, we should look at both low and high tails as well as averages, and perhaps individuals within the tails (Baye and Monseur 2016). The points raised in this literature review will be taken forward to our Findings and Discussion section.

Background – the WeeBee engine

Our engine is fundamentally linguistic in nature; children code narratives or stories by writing lines of Java code. The engine design is based on Systemic Functional Grammar (SFG) (Halliday 2004), where the engine's code statements used by children are closely derived from natural language clauses. Details of how this is achieved can be found in Price and Price-Mohr (2018a), the intention was to establish resonance between code statements and natural story language.

Figure 1 shows how the engine is presented to the children; on the left is a code-entry screen where they type their code, on the right is a canvas where their animation will appear, the toolbar has a few controls such as run, stop, file load, and save. The story presented here is from child D4.

Children typically add lines of code or modify previously added lines to produce an animation that captures their story in mind. Following the addition of code, they press the run button which compiles and executes their code to this point; also, a log file is written capturing their code, and

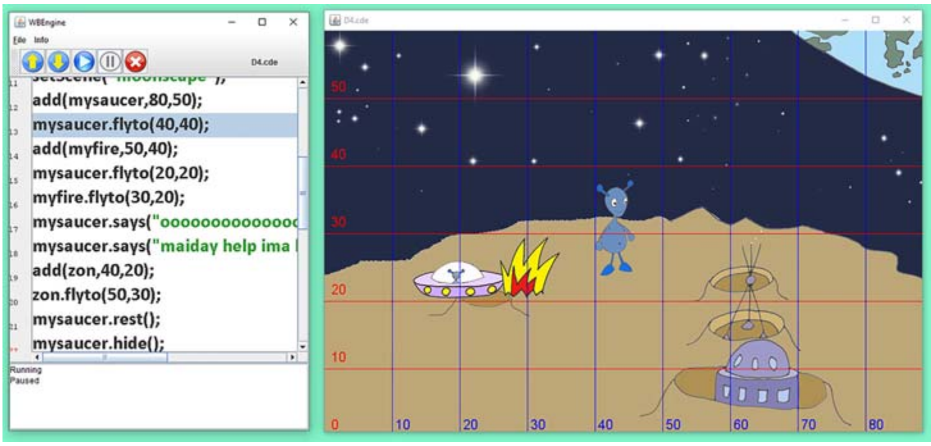


Figure 1. The WeeBee engine. On the left is the code-entry box where children type their Java code. When the run button is pressed, the animation of the code statements appears on the canvas at the right. The Cartesian grid helping placement of objects can be toggled on and off.

any compiler or run-time errors. This allows investigation of the children's *process* of coding. Indeed, children are learning about the process of programming, which has often been neglected; as Caspersen (2018, 11) notes 'Essentially programming is one of the best-kept secrets of programming education!'

The WeeBee engine assets (characters, props, scenery, and backgrounds) were designed explicitly to be gender neutral. These were imported from previous research into primary school literacy (Price-Mohr 2016; Price-Mohr and Price 2019). This research emphasised the need to create assets that make no reference to popular culture or to children's lives; the characters are animal-like but imaginary. These features ensure cultural and gender neutrality.

Research methodology

In this section, we discuss how the study was implemented, the participants, and ethical considerations. We introduce three measures used to evaluate the children's work. As detailed below, some measures evaluate the final product (entire program and animation quality); these may be thought of as measures of 'attainment'. Other measures are related to the development of the final product, i.e. the *process* of coding; these may be thought of as measures of 'progress'. Progress and Attainment are key features used by the school inspection system in England, applied especially to reading, writing, and mathematics (Ofsted 2019). We employed a mixed-methods approach; statistical analysis was applied to 'code coverage' (the choice of code methods used) and to the process of coding. The final animated stories were coded according to a rubric we developed, based on the literature, and this coding was then subject to statistical analysis.

Participants

Participants were drawn from an urban primary school class of grade 5 (ages 10–11). There were 17 girls and 15 boys arranged into 8 pairs of girls, 7 pairs of boys and one mixed pair. The latter was not included in this study. Four learning sessions each 45 min were delivered by the classroom teacher, who had received one hour of instruction from the research team. The children were told that by the end of their last session they should have completed a short narrative or story animation. The class had no prior experience of the WeeBee engine; we were therefore investigating the initial phase of learning to code.

Ethics

Ethical approval was granted by the University of Worcester. We followed the BERA Ethical Guidelines for Educational Research (2018). Voluntary informed consent was provided by the Headteacher, acting in *loco parentis*. Yet we felt it important that the children understood why the research was being conducted, what we hoped to achieve, and that they felt they were valued participants in the project. A discussion around these issues was led by the class teacher. The children (and the class teacher) were aware that any child could withdraw from the study; no child asked to withdraw. Children remained anonymous to the researchers; all data or information was communicated to the researchers using code names chosen by the children.

Code coverage measure

The engine provides a number of 'classes' of actions that each story character can use. These classes, inspired by SFG are summarised in Table 1 and capture a range of important SFG grammatical 'processes' or action words, (Halliday 2004, chapter 5). Children's final programs were analysed automatically, summing up the number of methods used for each class of method. This measure reflects the child's *attainment* in learning the engine's methods.

Since the raw data were typed ordinal, we used cross-tabulation with the Chi-Square test as the statistical tool.

The process of coding measure

Each time the children press the engine's 'run' button, a time-stamped log is created containing their code at that time, together with any compiler errors reported. Analysis of these logs provides information on their *process* of coding and therefore their progress. To capture the cognitive effort involved in creating meaningful code, we have previously defined the 'Purposeful Coding Effort' (PCE) measure (Price and Price-Mohr 2018b). This score reflects the addition, modification, and deletion of code line with clear intent; a score 1.0 and above reflects purposeful coding.

We report two run-based (progress) measures; errors/run and PCE/run; division by the total number of runs gives a measure which is relative to children's activity (coding then hitting the 'run' button). We also report two line-based (attainment measures) errors/line and PCE/line; division by the total number of lines in their program gives a measure which is relative to their final complete program, and not how they got there. In addition, we report children's activity as runs/min and lines/min. Again, these measure different things; a large value of runs/min would indicate sustained activity, a large value of lines/min could result from burst of activity punctuated by pauses. Finally, we classify the types of errors produced and subject this to analysis. Since the raw data were non-normal according to the Shaprio-Wilkes test we used the Wilkinson Sum Rank test for significance and effect size.

Table 1. Classes of methods available in the code with a small number of examples.

Class	Example methods	Description
move-to	pip.flyto(30,20); pip.runto(40);	Movement to a location in space
move-at	pip.jump(10); pip.spin();	Movement at the current location
transformative	pip.grow(1.5); pip.hide();	Change of character appearance
verbal	pip.says('Hello Grog');	Direct speech
cognitive	pip.thinks('He looks scary');	Expresses thoughts
relational	pip.feels(unhappy)	Expresses feelings or emotions

Note: Text quoted in verbal and cognitive methods appears on the screen in different colours at a unique place for each character.

Animation quality measure

The quality of the children's animations was scored against a rubric which was derived mainly from standard narratology research; details of the rubric and how it was developed, based on the literature, are presented below. Broadly speaking, we looked for chains of events with an explicit purpose or reason, other event chains where animacy could be perceived, and established characteristics of narrative or story such as the classical Proppian structure or 'canonicity and breach' (Bruner 1991). In addition, we assessed the spatial composition of the scene (and any scene changes), and how any character dialogue contributed to the narrative. We also measured 'low-level' attributes such as character movement and character/object interactions. Our final eight rubric categories are shown in Table 2. The animations were scored by two trained research assistants, inter-rater reliability was measured using Cohen's Kappa for ordinal data with kappa = 0.92, $p < .001$.

The fundamental backbone of a narrative is a chain of events in time (Genette 1982), and these may happen with various frequencies and durations and even out of order (category M). Yet Onega and Landa (1996) developed the concept of an event chain, asserting that *causality* must be present, to change a raw sequence of events into a story (category A). The relationship between events and how characters respond to them, displaying intelligence has been highlighted by Ryan (2007), this is included in category M. But a narrative is much more than a causal chain of events with character response. Following on from the work of Vladimir Propp (1968), Todorov (1968) argued that a narrative should follow a particular trajectory, from an initial equilibrium state which is disrupted into a state of disequilibrium, and at the end, equilibrium is restored (category S). All of this emphasises the role of time in the narrative. Gaps in perceived time can lead to suspense or surprise which according to Sternberg (2003) is an essential ingredient of a narrative (category H). Whilst a narrative unfolds in time, objects in the narratives are situated in space. Zoran (1984) presents us with a theory of narrative space and indicates that the narrative text provides a structure to the abstract narrative space. In our case, space is explicitly defined by the background and scenery rather than implicitly by reading text (category C). Diverging from narratology, it is important to consider the power of *perceived animacy*, where inanimate objects (such as elements of scenery) appear to be alive under simple

Table 2. Rubric for scoring the quality of the children's running animations.

	Animation features	Score
A	Object ***moves-to*** meaningful point	1
	Object moves with a purpose or cause	2
	Object appears or disappears with reason	1
	Object ***moves-at*** with reason	1
	Object changes appearance with reason	1
B	One object meets a second. The second does not move.	1
	One object moves away from a second which does not move	1
	Two objects meet, both move	2
	Two objects move apart	2
C	There are one or more scene changes	1
	Good composition of the scene	1–3
N	Object appears without reason	−1
	Object does ***move-at*** without reason	−1
	Object moves to no meaningful place	−1
M	There is a 'chain' of events in time	2*
	There is a 'chain' of events which has a goal.	3*
	Characters respond to events	2
V	Spoken/thought words related to actions	2*
	Spoken/thought words have a purpose	3*
H	There is some 'surprise' in the animation	2
	The reader could anticipate what comes next	2
S	There is the start of a story	1–2*
	There is a story with beginning, middle and end.	3–5*

Note: Reading down the table the categories move from animation events to story quality. Asterisks identify choices within the category.

conditions such as coming together, chasing each other, or simply by changing directions under certain conditions (van Buren, Gao, and Scholl 2017), and the motion of even a single object may be sufficient to impart animacy (Tremoulet and Feldman 2000). This forms category B.

Findings and analysis

In this section, we return to the three measures introduced above and summarise our findings using the appropriate statistical tools. We set these alongside key points identified in the literature, discussed in the Background section. To summarise our findings, we find no statistically significant differences in the process of coding, nor in the quality of animations produced. Some significant differences do emerge in the choice of the coding methods used ('code coverage').

Code coverage

The results of the cross-tabulation analysis, performed in 'R' are shown graphically in Figure 2; corresponding numerical values are presented in Table 3. The analysis yielded an overall Chi-squared = 34.73, df = 5, $p < .0001$ indicating a significant difference between the use of process classes. Overall, the data indicate that the total coverage of processes for boys (49.51%) and for girls (50.49%) is about the same. Also, the majority of processes used are move-to (31.89%), verbal (27.31%) and move-at (24.38%). Girls make more use of move-at processes (62.77%) and this contributes to 30.31% of their process makeup. Girls also make more use of cognitive processes (85.48%), but this contributes to a small amount (5.51%) of their process makeup. This does not support the suggestion that girls are less interested in 'things versus people' (Su, Rounds, and Armstrong 2009). Boys make more use of transformational processes (68.62%), but this makes up only a moderate amount (13.49%) of their process makeup. These processes, that change the appearance of characters, may follow on from boys' sensation-seeking tendency (Cross, Copping, and Campbell 2011). There is a slight, but not significant, difference in the use of verbal processes (in the favour of boys), move-to processes (in the favour of boys) and relational processes (in the favour of girls). It is interesting that boys used slightly

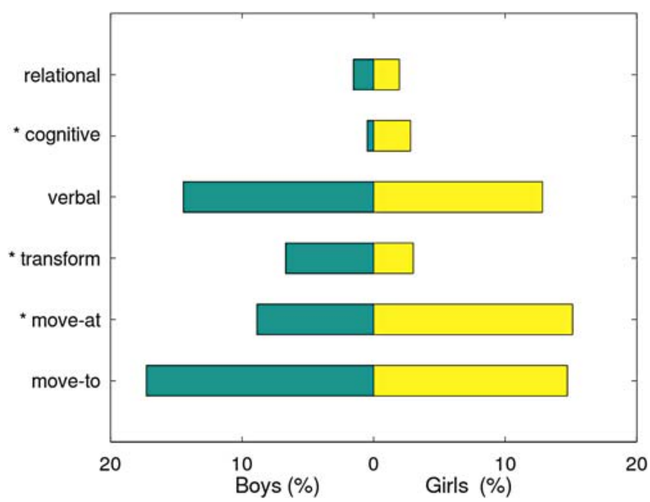


Figure 2. Code coverage: An overview of the percentage use of methods in each code class. The total length of each combined bar shows the percentage of methods from each class used by all children; the relative use is indicated by the bar split. Statistically significant differences are identified by asterisks. Relational and cognitive methods were used sparingly, but mostly by girls. The most prominent methods are move-to and verbal which show slightly more use by boys and move-at methods used slightly more by girls.

Table 3. Numerical results of code coverage analysis.

		move-to	move-at	trans-form	verbal	cognitive	relational	Row Total
Boys	Count	120	63	46	100	3	10	345
	Expected	110.00	84.36	33.67	94.59	11.27	12.19	
	Row %	34.95%	18.33%	13.49%	29.17%	0.96%	3.10%	49.51%
	Col %	54.42%	37.23%	<u>68.62%</u>	52.88%	14.52%	43.54%	
	Std. Res.	1.04	<u>−2.28</u>	<u>2.24</u>	0.66	<u>−2.37</u>	−0.42	
Girls	Count	101	106	21	89	19	13	352
	Expected	112.20	86.04	34.35	96.38	11.49	12.43	
	Row %	28.70%	30.31%	6.05%	25.49%	5.51%	3.94%	50.49%
	Col %	45.58%	62.77%	31.38%	47.12%	<u>85.48%</u>	56.46%	
	Std. Res.	1.03	<u>2.26</u>	<u>−2.22</u>	−0.656	<u>2.35</u>	0.42	
Col Total		222	170	68	190	22	24	698
		31.89%	24.38%	9.73%	27.31%	3.26%	3.52%	

Note: Columns label the process class; major row sections indicate boys and girls. Labels indicate for each section: Observed count, Expected count, Row% (per cent contribution of each process class to the total process makeup), Col % (per cent contribution of boys or girls to each process class), Std. Res. (the 'standard residual', which indicates significance of each entry). Those classes which are significant at $p = .05$ or better, are indicated by residual value magnitudes > 1.96, and are underlined.

more verbal processes which do not agree with the reported lower verbal fluency of boys (Hyde and Linn 1988).

The process of coding

The results of the process analysis are shown graphically in Figure 3; corresponding numerical values are presented in Table 4. There are no significant differences here, and indeed most differences are small.

Line-based measures indicate the nature of the final program and could be viewed as measures of attainment. Looking at the time required to complete the final code there is little difference in lines per minute with boys ($Mdn = 0.44$) compared with girls ($Mdn = 0.46$, $p = .87$, $r = 0.03$). Both groups worked overall at the same rate. The PCE per line for boys ($Mdn = 1.3$) showed little difference with girls ($Mdn = 1.3$, $p = .96$, $r = 0.01$) indicating both groups spent the same time on purposeful coding. Girls had to deal with slightly more errors than boys with error per line ($Mdn = 0.31$)

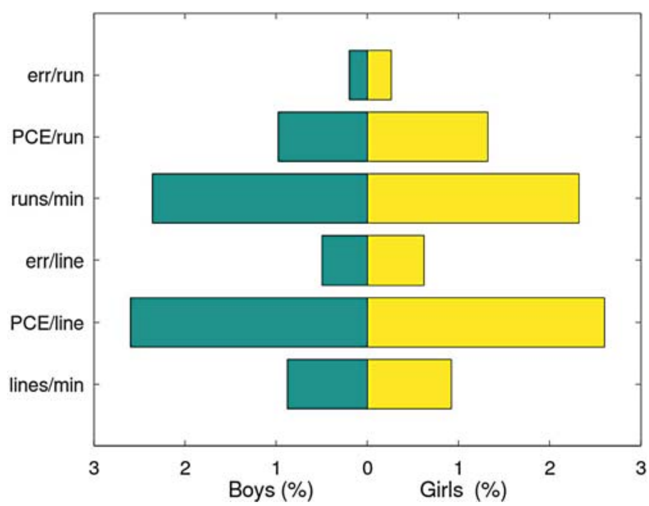


Figure 3. An overview of the process of coding. The total length of each combined bar shows the percentage of methods from each class used by all children; the relative use is indicated by the bar split. There are no statistically significant differences; boys and girls perform almost identically, with girls appearing slightly more error-prone.

Table 4. Numerical results of the Process of Coding analysis (Wilcoxon Sum Rank test).

Measure	Median (Boys)	Mean (Boys)	Median (Girls)	Mean (Girls)	W	*p*	*r*
Errors per run	0.1	0.11	0.13	0.14	60	.09	−0.31
PCE per run	0.49	0.64	0.66	0.66	89.5	.78	−0.05
Runs per minute	1.18	1.36	0.89	1.16	112.5	.46	−0.14
Errors per line	0.25	0.36	0.31	0.42	88	.73	−0.06
PCE per line	1.3	1.6	1.3	1.4	97.5	.96	−0.01
Lines per minute	0.44	0.48	0.46	0.45	100	.87	−0.03

Note: There are no significant differences at the $p < .05$ level, though the errors per run is significant at the $p < .1$ level with small to medium effect size.

compared with boys (*Mdn* = 0.25, $p = .73$, $r = 0.06$). Taken together, these results suggest there are no gender differences in attainment; both boys and girls are able to debug and complete meaningful programs, also that there is no gender difference in progress; on average boys and girls are equally adept at learning to code. They are all able to correct errors, which is not a simple skill even for college students. This may seem surprising, since a gender difference (in favour of girls) in writing productivity and writing quality has been identified (Adams and Simmons 2019) which has evidently not appeared in composition using the medium of code.

Run-based measures indicate activity during the learning to program phase and could be viewed as measures of progress. There is little difference in programming activity, runs per minute with boys (*Mdn = 1.18*) compared with girls (*Mdn* = 0.89, $p = .46$, $r = 0.14$). Girls have a slightly larger PCE/run (*Mdn* = 0.66) compared with boys (*Mdn* = 0.49, $p = .78$, $r = 0.05$) suggesting they are thinking more. Looking in detail at changes made to existing code, around 55% of these were made to the add() method in order to tweak the positioning of scenery and characters to obtain the desired scene layout, and there was no difference between boys and girls. This suggests that boys gave as much thought to their work as girls and does not confirm the notion that boys 'just get on with it' (Funke et al. 2015; Papavlasopoulou, Sharma, and Giannakos 2020). There was one exception, a boy (D11) who experimented with almost random object sizes without any purpose; his final animation was of very low quality. There were differences in other code changes made; girls were reluctant to delete lines of working code while boys often deleted lines. Girls made more changes to methods, such as changing walking to flying motion, which may suggest girls had a more comprehensive understanding of the methods available.

Looking at the errors made per run, there is little difference between errors per run with boys (*Mdn* = 0.10) compared with girls (*Mdn* = 0.13, $p = .09$, $r = 0.31$) with medium effect size. The types of errors made were broadly the same; however, boys made more typographical errors and girls more syntax errors (incorrect code statement). Three boys showed evidence of error correction by experimentation rather than by logical thought (D7, D11, D10).

Animation quality

The results of this analysis are shown in graphically in Figure 4; corresponding numerical values are presented in Table 5. There are no significant differences. Around 53% of the total marks are assigned to girls suggesting they produced slightly better animations. Looking across the categories, girls slightly outperformed boys in each category except verbal (speech). This is interesting since verbal fluency is often associated with girls (Hyde and Linn 1988). Girls scored higher on high-level measures of narrative (56.8% for story structure and 58.68% for scene composition), the slight difference perhaps reflecting girls' greater interest in reading and writing stories (rather than e.g. instructional texts).

The narrative conveyed by the animations showed no significant gender difference 'things versus people' (Su, Rounds, and Armstrong 2009); there was a slight, but not significant, tendency for girls to create more people-based animations. Three of the boy's animations involved the use of fire to

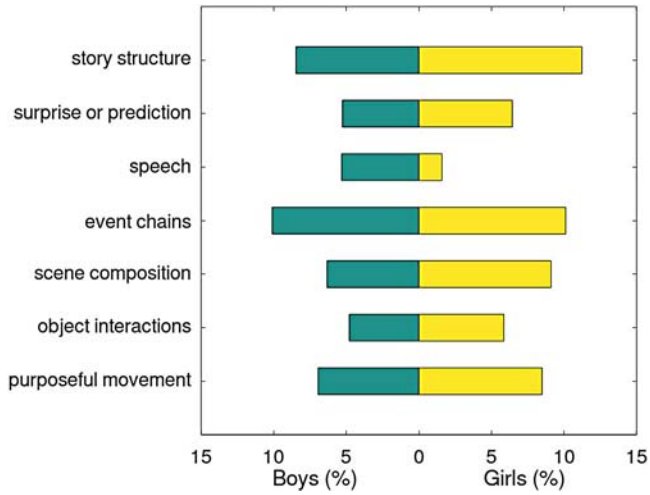


Figure 4. Results of the animation quality analysis. The total length of each combined bar shows the percentage of methods from each class used by all children; the relative use is indicated by the bar split. While there are no statistically significant differences, boys score higher on the use of speech and girls score better on scene composition and story structure.

convey danger or destruction, perhaps an indication of boys' attention seeking (Cross, Copping, and Campbell 2011).

Discussion

This study has explored potential gender differences in coding stories using the WeeBee engine, within the context of gender differences reported in the literature. Overall, we find no strong evidence for gender differences in coding strategies or quality of animations, which agrees with Hyde's gender equality hypothesis (Hyde 2005). We find little or no support for the greater male variability hypothesis (Feingold 1992; Machin and Pekkarinen 2008).

Turning to other details reported in the literature, we find slight evidence that girls prefer 'people versus things' (Su, Rounds, and Armstrong 2009), and that boys are more sensation seeking (Cross, Copping, and Campbell 2011). Perhaps most surprising is that boys do not seem averse to using verbal processes, which is confirmed by the animation quality analysis; previous research has suggested that girls are more verbally fluent (Hyde and Linn 1988).

Table 5. Numerical results of animation quality analysis.

		A	B	C	M	V	H	S	Row Total
Boys	Count	13	9	12	19	10	10	16	89
	Expected	13.73	9.47	13.73	17.99	6.15	10.	17.52	
	Row %	14.61%	10.11%	13.48%	21.35%	11.24%	11.24%	17.98%	47.34%
	Col %	44.83%	45.00%	41.38%	50.00%	76.92%	45.45%	43.24%	
	Std. Res.	−0.20	−0.15	−0.47	0.24	1.55	−0.13	−0.36	
Girls	Count	16	11	17	19	3	12	21	99
	Expected	15.27	10.53	15.27	20.01	6.85	11.59	19.48	
	Row %	16.16%	11.11%	17.17%	19.19%	3.03%	12.12%	21.21%	52.66%
	Col %	55.17%	55.00%	58.62%	50.00%	23.08%	54.55%	56.76%	
	Std. Res.	0.19	0.14	0.44	−0.23	−1.47	0.12	0.34	
Col Total		29	20	29	38	13	22	37	188
		15.43%	10.64%	15.43%	20.21%	6.91%	11.70%	19.68%	

Note: Categories are A, movement with purpose; B, object interactions; C, scene composition; M, event chains; V, speech; H, surprise, or prediction; S, story structure. There are no significant differences, but larger differences are underlined.

During the process of coding, we find there is no gender difference in productivity or quality of writing (coding) which again is surprising given the results reported by Adams and Simmons (2019). Specifically, there is no difference in the effectiveness of girl pairs over boy pairs, which again challenges the findings reported by Tsan, Boyer, and Lynch (2016). All children were able to create a working program, unlike the Scratch coding reported by (Funke and Geldreich 2017). Looking at the code composition, and process of coding, refutes the idea that girls think first while boys just get on with it (Funke et al. 2015; Papavlasopoulou, Sharma, and Giannakos 2020).

Turning now to variation, comparison of the various line-and-run-based measures does show large variations in the scores around the mean, but these are about the same for boys and girls, suggesting a class of mixed ability. For example, taking the percentage runs to correct errors we find the mean for boys is 40.5% (St. Dev 6.6%) and for the girls 41.7% (St. Dev. 6.9%). This does not agree with the male variability hypothesis (Feingold 1992; Machin and Pekkarinen 2008), where we would expect a larger variation for boys.

Following the suggestions of Baye and Monseur (2016), we looked for children in both the low and high tails according to progress (lines/min and PCE/run) and attainment (errors/line and animation quality). Children were assigned to a tail when all four of these measures agreed; the low tail comprised D7, D11 (boys) and D9 (girl) and the upper tails D4 (boy) and D14 (girl). Both D7 and D11 made a large number of errors, and multiple errors of the same type, suggesting impatience or lack of learning. Child D9 made few errors, but took many runs to correct these, as did D7. Both of these children produced very short final programs and correspondingly poor animations. As mentioned, D11 attempted to correct errors by experimentation, and he also experimented with parameters assigning 'silly' (very large or small) values. We have experienced this before working with children in the classroom.

In the top tail, D1 (girl) and D10 (boy) made very few errors and constructed programs at a high speed of lines per minute. Their animations were outstanding[1] D1 crafted a story of two friends playing and discovering a magic carpet which transported them to the Moon. Realising they are stuck, they ask Drax (the friendly dragon) to help them return. Drax bundles them into a spaceship which takes them back to Earth. The story makes good use of speech, scene changes, and suitable action, and forms a complete story. D10 tells the start of a story where Drax (here unfriendly) destroys Grog's favourite tree with a ball of fire. Grog and Saff then start to plan how to take revenge. Animation D11 paints a different picture. Here there is no storyline; the animation consists of multiple scene-changes often with solid colour background. Characters undergo meaningless stretching transformations, and speech is disconnected. There is a total lack of coherence.

Conclusion

In conclusion, we report no meaningful gender differences in coding using the WeeBee engine; moreover, we suggest that gender differences, in general, may not be as fundamental as previously believed, and we agree with Hyde (2014) that other factors may have been at work in many studies where moderate or large effect sizes were reported. Recent research suggests that there are little cognitive, neurophysiological, or neuroanatomical underpinnings of such differences (Jancke 2018). Also, there is evidence that any difference may be exaggerated by poor instruction (Price-Mohr and Price 2017) or by no instruction at all (Hyde 2014).

What are the implications of our findings for the teacher-educator or classroom practitioner? First, we agree with Hyde (2014) that educators should not assume that gender differences exist in this learning context, and therefore should not assume a different approach should be taken in teaching boys and girls, as was the case for Storytelling Alice, Kelleher, Pausch, and Kiesler (2007). Concerning learning to code, we have demonstrated that our approach using the WeeBee engine to create stories realised through animations is *fundamentally* gender neutral. We trace this to the design of the engine which focussed on gender neutrality, and lack of reference to popular culture, in order to appeal to a wide variety of personalities. This makes it a useful tool in learning to code since teachers do not need to give explicit attention to gender. In addition to our findings about gender difference in coding, we hope that the approaches to analysis we have presented may be taken forward as 'tools' by other researchers, in particular code coverage, the measures of coding progress (PCE), and the animation rubric. Finally, we have found confirmatory evidence that young children are able to code in text using a professional language (Java) and we encourage educators to adapt our approach as part of their Computing delivery. The cross-curriculum benefits are clear; we have emphasised the links with Literacy, links to creative art could be made, and also to mathematics. To this end, the most recent version of our engine provides 'Turtle Graphics' to explicitly help the learning of geometrical concepts. The WeeBee engine and supporting resources are freely available from the corresponding author.

Note

1. See supplemental material online, animations for D1, D10 and D11.

Acknowledgments

We would like to thank the class teacher and the children of the participating school, especially child D4 who gave permission to include their code in this article.

Disclosure statement

No potential conflict of interest was reported by the author(s).

ORCID

Colin B. Price http://orcid.org/0000-0002-2173-9897
Ruth Price-Mohr http://orcid.org/0000-0001-9494-6398

References

Adams, A.-M., and F. R. Simmons. 2019. "Exploring Individual and Gender Differences in Early Writing Performance." *Reading and Writing* 32: 235–263.

Arfé, B., T. Vardanega, C. Montuori, and M. Lavanga. 2019. "Coding in Primary Grades Boosts Children's Executive Functions." *Frontiers in Psychology* 10: 2713. doi:10.3389/fpsyg.2019.02713.

Baye, A., and C. Monseur. 2016. "Gender Differences in Variability and Extreme Scores in an International Context." *Large-scale Assessments in Education* 4: 1.

Bruner, J. 1991. "The Narrative Construction of Reality." *Critical Inquiry* 18: 1–21.

Buechley, l., M. Eisenberg, J. Catchen, and A. Crockett. 2008. "The LilyPad Arduino: Using Computational Textiles to Investigate Engagement, Aesthetics, and Diversity in Computer Science Education". Paper presented at the Proceedings of the SICCHI Conference on Human Factors in Computing Systems, Florence, Italy April 5–10.

Burnett, C. 2016. *The Digital Age and Its Implications for Learning and Teaching in the Primary School*. York: Cambridge Primary Review Trust.

Çakir, N. A., A. Gass, A. Foster, and F. J. Lee. 2017. "Development of a Game-Design Workshop to Promote Young Girl's Interest Towards Computing Through Identity Exploration." *Computers & Education* 108: 115–130. doi:10.1016/j.compedu.2017.02.002.

CAS (Computing at School). 2015. "QuickStart Computing." https://community.computingatschool.org.uk/resources/3042/single#v1.

Caspersen, M. E. 2018. "Teaching Programming." In *Computer Science Education: Perspectives on Teaching and Learning in School*, edited by Computer Science, S. Sentance, E. Barendsen, and C. Schulte, 109–130. London: Bloomsbury Academic.

Cross, C. P., L. T. Copping, and A. Campbell. 2011. "Sex Differences in Impulsivity: A Meta-Analysis." *Psychological Bulletin* 137: 97–130.

Denner, J., and S. Campe. 2018. "Equity and Inclusion in Computer Science Education." In *Computer Science Education: Perspectives on Teaching and Learning in School*, edited by S. Sentance, E. Barendsen, and C. Schulte, 189–205. London: Bloomsbury Academic.

Driessen, G., and A. van Langen. 2013. "Gender Differences in Primary and Secondary Education: Are Girls Really Outperforming Boys?" *International Review of Education* 59: 67–86. doi:10.1007/s11159-013-9352-6.

Else-Quest, N. M., J. S. Hyde, and M. C. Linn. 2010. "Cross-National Patterns of Gender Differences in Mathematics: A Meta-Analysis." *Psychological Bulletin* 136 (1): 103–127. doi:10.1037/a0018053.

Feingold, A. 1992. "Sex Differences in Variability in Intellectual Abilities: A New Look at an Old Controversy." *Review of Educational Research* 62 (1): 61–84.

Feingold, A. 1994. "Gender Differences in Personality: A Meta-Analysis." *Psychological Bulletin* 116: 429–456.

Funke, A., M. Berges, A. Muhling, and P. Hubwieser. 2015. "Gender Differences in Programming: Research Results and Teachers' Perception." Koli Calling '15, November 19–22, Koli, Finland.

Funke, A., and K. Geldreich. 2017. "Gender Differences in Scratch Programs of Primary School Children". Processings of WiPSCE '17, Nijmegen, Netherlands.

Genette, G. 1982. *Figures of Literary Discourse*. Oxford: Blackwell.

Halliday, M. A. K. 2004. *Halliday's Introduction to Functional Grammar*. Oxford: Routledge.

Husbye, N. E., B. A. Buchholz, L. S. Coggin, C. Wessel-Powell, and K. E. Wohlwend. 2012. "Critical Lessons and Playful Literacies: Digital Media in PK-2 Classrooms." *Language Arts* 90 (2): 82–92. http://www.jstor.org/stable/41804380.

Hyde, J. S. 2005. "The Gender Similarities Hypothesis." *Merican Psychologist* 60 (6): 581–592.

Hyde, J. S. 2014. "Gender Similarities and Differences." *Annual Review of Psychology* 65: 373–398. doi:10.1146/annurev-psych-010213-115057.

Hyde, J. S., and M. C. Linn. 1988. "Gender Differences in Verbal Ability: A Meta-Analysis." *Psychological Bulletin* 104: 53–69.

Jancke, L. 2018. "Sex/Gender Difference in Cognition, Neurophysiology, and Neuroanatomy." *F1000Research* 7: 805.

Kalelioglu, F. 2015. "A New Way of Teaching Programming Skills to K-12 Students." *Computers in Human Behaviour* 52: 200–210. doi:10.1016/j.chb.2015.05.047.

Kelleher, C., R. Pausch, and S. Kiesler. 2007. "Storytelling Alice Motivates Middle School Girls to Learn Computer Programming". Proceesdings of the SIGCHI Conference on Human Factors in Computing Systems, San Jose California USA 28 April 2007–3 May 2007, 1455–1464.

Machin, S., and T. Pekkarinen. 2008. "Global Sex Differences in Test Score Variability." Science 322 (5906): 1331–1332.

Miller, J. 2019. "STEM Education in the Primary Years to Support Mathematical Thinking: Using Coding to Identify Mathematical Structures and Patterns." *ZDM* 51: 915–927.

O'Dea, R. E., M. Lagisz, M. D. Jennions, and S. Nakagawa. 2018. "Gender Differences in Individual Variation in Academic Grades Fail to Fit Expected Patterns for STEM." *Nature Communications* 9: 3777.

Ofsted. 2019. "School Inspection Update." https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/772056/School_inspection_update_-_January_2019_Special_Edition_180119.pdf.

Onega, S., and J. A. G. Landa. 1996. *Narratology: An Introduction*. London: Longman.

Papavlasopoulou, S., K. Sharma, and M. N. Giannakos. 2020. "Coding Activities for Children: Coupling Eye-Tracking with Qualitative Data to Investigate Gender Differences." *Computers in Human Behaviour* 105. https://doi.org/10.1016/j.chb.2019.03.003.

Price, C. B., and R. M. Price-Mohr. 2018a. "Stories Children Write When Coding: A Cross-Disciplinary Approach for the Primary Classroom." *Cambridge Journal of Education* 48 (6): 735–747. doi:10.1080/0305764x.2017.1418834.

Price, C. B., and R. M. Price-Mohr. 2018b. "An Evaluation of Primary School Children Using a Text-Based Language (Java)." *Computers in the Schools* 35 (4): 284–301. doi:10.1080/07380569.2018.1531613.

Price-Mohr, R. M. 2016. *Comparing a Controlled Levelled Text with a Language Rich Text in a Beginner Reading Scheme*. York: University of York. http://etheses.whiterose.ac.uk/12904/.

Price-Mohr, R. M., and C. B. Price. 2017. "Gender Differences in Early Reading Strategies: A Comparison of Synthetic Phonics Only with a Mixed Approach to Teaching Reading to 4-5 Year Old Children." *Early Childhood Education Journal* 45: 613–620. doi:10.1007/s10643-016-0813-y.

Price-Mohr, R. M., and C. B. Price. 2019. "A Comparison of Children Aged 4-5 Years Learning to Read Through Instructional Texts Containing Either a High or Low Proportion of Phonically Decodable Words." *Early Childhood Education Journal* 48: 39–47. doi:10.1007/s10643-019-00970-4.

Propp, V. 1968. *Morphology of the Folktale*. Austin: University of Texas Press.

Reilly, D. 2015. "Gender Differences in Reading from a Cross-cultural Perspective – The Contribution of Gender Equality". Paper presented at the International Convention of Psychological Science, Amsterdam, Netherlands.

Resnick, M., J. Maloney, A. Monroy-Hernandez, N. Rusk, E. Eastmond, K. Brennan, A. Millner, et al. 2009. "Scratch: Programming For All." *Communications of the ACM* 51 (11): 60–67.

Ryan, M. L. 2007. ""Toward a Definition of Narrative"." In *The Cambridge Companion to Narrative*, edited by D. Herman, 22–35. Cambridge: Cambridge University Press.

Sternberg, M. 2003. "Universals of Narrative and Their Cognitivist Fortunes (I)." *Poetics Today* 24 (2): 326–328.

Su, R., R. Rounds, and P. Armstrong. 2009. "Men and Things, Women and People: A Meta-Analysis of Sex Differences in Interests." *Psychological Bulletin* 135: 859–884.

Todorov, T. 1968. "La Grammaire du Recit." *Languages* 12: 94–102.

Tremoulet, P. D., and J. Feldman. 2000. "Perception of Animacy from the Motion of a Single Object." *Perception* 29: 943–951.

Tsan, J., K. E. Boyer, and C. F. Lynch. 2016. "How Early Does the CS Gender Gap Emerge? A Study of Collaborative Problem Solving in 5th Grade Computer Science". SIGCSE '16 Menphis, Tennessee, USA.

van Buren, B., T. Gao, and B. J. Scholl. 2017. "What are the Underlying Units of Perceived Animacy? Chasing Detection is Intrinsically Object-Based." *Psychological Bulletin Review* 24: 1604–1610.

Walker, S., and D. Berthelsen. 2017. "Gender Differences in Early Literacy and Mathematics Achievement and Self-Regulatory Behaviours in the First Year of School: An Australian Study." *Australasian Journal of Early Childhood* 42 (1): 70–78.

Wang, M., and J. L. Degol. 2017. "Gender Gap in Science, Technology, Engineering, and Mathematics (STEM): Current Knowledge, Implications for Practice, Policy, and Future Directions." *Educational Psychology Review* 29: 119–140. doi:10.1007/s10648-015-9355-x.

Wohlwend, K. E. 2015. "One Screen, Many Fingers: Young Children's Collaborative Literacy Play With Digital Puppetry Apps and Touch Screen Technologies." *Theory Into Practice* 54 (2): 154–162. doi:10.1080/00405841.2015.1010837.

Zoran, G. 1984. "The Construction of Reality in Fiction." *Poetics Today* 5 (2): 309–335.