International Journal of Child-Computer Interaction 29 (2021) 100293

Contents lists available at ScienceDirect

International Journal of Child-Computer Interaction

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

Research paper

Interaction effects of race and gender in elementary CS attitudes: A validation and cross-sectional study

Jessica Vandenberg [a,*], Arif Rachmatullah [a], Collin Lynch [a], Kristy E. Boyer [b], Eric Wiebe [a]

[a] North Carolina State University, USA
[b] University of Florida, USA

ARTICLE INFO

Article history:
Received 8 January 2021
Received in revised form 16 March 2021
Accepted 19 March 2021
Available online 6 April 2021

Keywords:
Computer Science Attitudes
Invariance test
Scale validation
Primary school

ABSTRACT

Computer science (CS) initiatives for elementary students, including brief Hour of Code activities and longer in- and after-school programs that emphasize robotics and coding, have continued to increase in popularity. Many of these initiatives are intended to increase CS exposure to students who historically have been underrepresented in CS academic trajectories, including women and students of color. This study aimed at examining the gender and race difference in elementary students' attitudes toward CS. To that end we developed and validated a survey instrument called Elementary Computer Science Attitudes (E-CSA) which consisted of the constructs of CS self-efficacy and outcome expectancy, through a combination of classical test theory and item response theory. The target audience for this instrument and study was upper elementary students (grades 4 and 5, ages 8 to 11). The E-CSA was found to be a gender and race bias-free instrument. A two-way ANOVA test was then used to answer research questions. We found no significant interaction effect between gender and race in the two constructs of CS Attitudes. We also did not see a significant difference based on race. However, a significant difference was found in both CS attitudes constructs based on gender, whereby male students had higher CS attitudes than female students. We discuss our findings from the perspective of the equity issue in CS education. Furthermore, we believe the E-CSA instrument can inform classroom-based interventions, the development of curricular materials, and reinforce findings from cross-sectional CS studies.

1. Introduction

Despite a need for computing knowledge for educational purposes and career advancement, there remain challenges to attracting students to computationally-intensive STEM fields and retaining them once there (Belser, Prescod, Daire, Dagley, & Young, 2017; Lent, Lopez, Lopez, & Sheu, 2008). Women and historically underrepresented minorities (URMs) in these fields are especially likely to not take computer science (CS) classes or apply to CS majors or to not persist once enrolled (Sax, Lehman et al., 2017; Sax, Zimmerman et al., 2017). Although elementary students are years away from having to declare a major or seek a job, this population is a critical point for learning foundational CS concepts and, perhaps more importantly, how CS practices can be a powerful way of approaching a learning task. Moreover, since positive affective orientation is critical to students maximizing the benefits of these activities, we need to be able gauge their interests, both proximal and distal (Yoo et al., 2017).

Early exposure to high quality computing experiences may inform a young person's trajectory toward a STEM career, although there are myriad social forces at play that adversely affect girls and students of color having a positive orientation toward either CS or STEM. In fact, children as young as six readily express gendered stereotypes such that boys are better at programming and robotics than girls (Master, Cheryan, Moscatelli, & Meltzoff, 2017). Internalization of these beliefs is particularly harmful to girls, as it affects their interest in and self-efficacy for these subjects. Many URMs, historically marginalized in this field, feel unwelcome in or disconnected from CS; a lower sense of belonging (Johnson, 2011; Leath & Chavous, 2018) and racial/ethnic stereotypes (Margolis, Estrella, Goode, Holme, & Nao, 2017) have been cited as reasons why.

A lack of validated attitudinal instruments at the elementary level hampers our ability to both study and address the issue of the gender and racial/ethnic gap in CS. Minimal validation research has been completed on students' CS attitudes that both adheres to core psychological theory and utilizes powerful psychometric analytic methods. Two major exceptions follow. Mason

* Corresponding author.
E-mail addresses: jvanden2@ncsu.edu (J. Vandenberg), arachma@ncsu.edu (A. Rachmatullah), cflynch@ncsu.edu (C. Lynch), keboyer@ufl.edu (K.E. Boyer), wiebe@ncsu.edu (E. Wiebe).

https://doi.org/10.1016/j.ijcci.2021.100293
2212-8689/

and Rich (2020) represents one of the few serious efforts in this area. They validated their Elementary Student Coding Attitudes Survey (ESCAS) – centered around concepts of coding confidence, interest and utility, in addition to social influence, and perception of coders – using confirmatory factor analysis (CFA) and structural equation modeling (SEM). Despite these efforts, they did not test if their instrument was psychometrically free from gender or race bias. Rachmatullah, Wiebe et al. (2020) validated a middle grades CS attitudes instrument centered around the concepts of self-efficacy and outcome expectancy by using the combination of classical test theory and item response theory Rasch techniques. This middle grades instrument was analyzed and found psychometrically free of gender and race bias, thus it is a robust starting point we use here for a new instrument to measure and investigate gender and race attitudes at the elementary level.

1.1. Theoretical framework

Bandura, Freeman, and Lightsey (1999) maintained that individuals are motivated by their beliefs in their capabilities to complete a task – called self-efficacy – and that completing that task will ultimately produce a desired outcome, called outcome expectancy. Pajares (1996) argued for task specificity in designing instruments and assessing students' self-efficacy; in our case, the specific task is coding within the domain of computer science. Further, we make use of expectancy-value theory (Eccles & Wigfield, 2002; Wigfield & Eccles, 2000) and its contention that students make purposeful academic decisions based on their expectations for success. In turn, these outcome expectancies influence a student's willingness to select and engage in a task, as well as to persist during challenge.

1.2. Related work

1.2.1. Self efficacy and outcome expectancy

Prior research on self-efficacy and outcome expectancy beliefs is varied in terms of gender differences. Older literature (e.g. Zeldin, Britner, & Pajares, 2008) has reported gendered differences in the sources of self-efficacy, with mastery experiences being the primary source of males' self-efficacy and females' relying more upon relational information (i.e., social input from peers, teachers, parents, and larger society) to inform their self-efficacy. Early practice and success continues to be a persistent factor. In fact, Lishinski, Yadav, Good, and Enbody (2016) found that self-efficacy predicted students' course outcomes (i.e., exam scores), but gender powerfully affected how students' self-efficacy changed in response to performance feedback early in the course; specifically, early failures or perceived setbacks may prompt female students to disengage from the CS course. It is therefore necessary for educators and researchers to explore this decrease in students' competence beliefs as it greatly affects students' academic performance. Muenks, Wigfield, and Eccles's review of expectancy and competence beliefs indicated that children's expectancy-related beliefs tend to decline from elementary through high school, although students follow different trajectories across different subject areas and these beliefs change based on performance. For elementary-aged female students, self-efficacy is significantly related to their CS career orientation (Aivaloglou & Hermans, 2019).

Research indicates that when students are exposed to negative gender-based stereotypes and they readily endorse those beliefs, their grades and career intentions are affected (Plante, De la Sablonnière, Aronson, & Théorêt, 2013). Master and Meltzoff's extensive review of the literature around stereotypes, STEM, gender, and motivation resulted in their development of a model that underscores the role stereotypes and students' beliefs, attitudes, and behaviors have on their interest and performance in STEM.

Google Inc. and Gallup Inc.'s report on diversity gaps in CS foregrounds Black/African–American students' higher confidence level and interest in CS compared to White and Hispanic students; this clearly supports the idea that confidence and interest are not the only factors that contribute to (under) representation in the field. James DiSalvo et al. (2011) reported that although Black/African–American males enjoyed playing video games, they often did not extend that interest to the computing concepts used to build the games. Findings such as these point to the need for instrument development and further research into the relationship of key demographic factors, beliefs, and outcomes regarding CS education.

1.2.2. Gender and ethnicity in CS

There is ample evidence to suggest that CS suffers from a lack of inclusivity. Women and girls often feel unwelcome (Beyer, 2014), suffer from lower confidence (Beyer, Rynes, Perrault, Hay, & Haller, 2003), or are downright excluded from CS courses and computing in general (Cheryan, Master, & Meltzoff, 2015; Cheryan, Plaut, Davies, & Steele, 2009). In a landmark study, Sax, Lehman et al. (2017) found several notable contributors to the gender gap in CS. In particular, they found that women self-reported lower math ability than male counterparts, held a social activist orientation, and felt less compelled to contribute to the scientific community. At the university level, some have found that when CS is taught using a pair programming approach, women perform better and persist in CS courses (Werner, Hanks, & McDowell, 2004) and self-report higher confidence than those required to work individually (McDowell, Werner, Bullock, & Fernald, 2006). Research at the middle school level (Buffum et al., 2015) and elementary level (Tsan, Boyer, & Lynch, 2016) indicates that gender differences are present in students' CS experiences, but how these manifest are quite different. Buffum et al. (2015) found that repeated exposure to CS concepts compensates for differences in students' prior computing experiences, whereas Tsan et al. (2016) found that girls' final CS products were significantly lower in quality compared to all boy groups and mixed gender groups.

Paralleling findings on gender, some research suggests that CS is not particularly open to a range of ethnicities and races. Underrepresented minorities (URMs) often encounter stereotypes about who is 'good' at CS (Margolis et al., 2017), and these stereotypical attributes tend to include high intelligence, limited social skills, and being white or Asian. Moreover, students tend to report that access and wealth positively affects one's ability to participate in CS and that wealth and access are often related to race and ethnicity. As a result, URM often are prevented from developing a sense of belonging in CS which then impedes their interest in pursuing additional coursework, a major, or a career in CS (Sax, Zimmerman et al., 2017). Of particular interest is how the intersection of race and gender might influence a student's CS trajectory; recent findings by Scott and colleagues (Scott, Martin, McAlear, & Koshy, 2017) highlight how female students of color had lower levels of engagement and interest, stating "...being a member of a marginalized gender group plays a unique role and has a multiplying (negative) effect" (Scott et al., 2017, p. 255). Also of note is that most of this literature focuses on older populations, yet we know (e.g. Aladé, Lauricella, Kumar, & Wartella, 2020; Mulvey & Irvin, 2018) that these social forces start affecting children at younger ages.

1.2.3. Other CS attitudes instruments

Our work has been informed by some notable prior research on CS attitudes. Kukul, Gökçearslan, and Günbatar's worked with 12 to 14 year old students in Turkey to produce the Computer Programming Self-Efficacy Scale (CPSES). This 31-item, unidimensional scale queries students on their self-efficacy for specific computing actions, such as "I know where to write the program codes". Self-efficacy, interest, and collaboration drove Kong, Chiu, and Lai (2018) to develop and validate a programming empowerment instrument for 4th through 6th grade students. Their 24-item instrument includes statements like "Programming is important to me" and "I like to program with others". More recently, and as noted earlier, Mason and Rich (2020) validated their Elementary Student Coding Attitudes Survey. This 23-item, 5-factor instrument queries 4th through 6th grade students on statements such as "Coding is interesting" and "Kids who code are smarter than average". All three of these instruments fall short of what is needed for evaluating young students' CS attitudes, despite analyzing children's responses around the same grade bands. They are all quite lengthy at 23 to 31 items and none of them evaluated if the instrument was free from bias. Moreover, the (Kukul et al., 2017) instrument has not been validated in English.

2. Current work

There is a clear need to research students' attitudes by race/ethnicity and gender, especially beginning at a young age. Per the review above, there is a lack of a brief, targeted, validated, and psychometrically bias-free instrument that measures CS Attitudes in upper elementary students and which accounts for the unique developmental differences of this population. To account for this, we detail below our qualitative procedures for ensuring young students understood our item wording (Vandenberg et al., 2020), after which we follow similar validation procedures as Rachmatullah, Wiebe et al. (2020). This instrument, the E-CSA, is then used to measure upper elementary (4th and 5th grade) students' attitudes toward computer science, with particular focus on the effect of race/ethnicity and gender on their responses. The following research questions guide this investigation:

(1) With regards to the validation of the instrument, what model best represents the dimensionality and internal structure of E-CSA?
(2) What is the relationship between elementary students' CS attitudes, measured on the E-CSA, and their CS conceptual understanding, measured on the E-CSCA (Elementary Computer Science Concepts Assessment)?
(3) What is the influence of race/ethnicity and gender on students' responses on the E-CSA instrument?

3. Methodology

3.1. Item development

The items for our instrument were based on the previously validated Engineering and Technology attitudes subscale of the Student Attitudes toward STEM (S-STEM) Survey (Friday Institute for Educational Innovation, 2012). The S-STEM survey has been used with over 15,000 4th through 12th grade US students (Wiebe, Unfried, & Faber, 2018). We then engaged in an iterative process of cognitive interviews with a diverse group of 98 4th and 5th grade students on their understanding of the items (Vandenberg et al., 2020). Findings from this process indicated that upper elementary aged students conceptualized of *doing* computer science as 'coding.' To make this lean instrument appropriate for young students, we privileged their words and the types of tasks in which they engaged in what we, as researchers and practitioners, consider computer science. As such, we used the word 'coding' because children were not able to define computer science. This rigorous process resulted in a final set of 11 Likert-scale items with 5 points from *strongly disagree* to *strongly agree* that reflected the modified wording of coding and computer science rather than engineering and technology. This instrument, the E-CSA, is based on two psychological constructs, self-efficacy (denoted as SE_) and outcome expectancy (denoted as OE_) (see Table 1).

3.2. Sample and contexts

Following university IRB approval, which required both parental consent and minor student assent, a total of 169 students consented/assented to take the E-CSA instrument as part of either a classroom-based study or a standalone survey administration for the purposes of this validation. This number is sufficient to perform IRT Rasch and obtain stable item calibrations and person measure estimates (Chen et al., 2014; Linacre, 1994). Participating students were third through fifth grade students (ages 8–11), with 5th grade students representing 66% of the sample and female students accounting for approximately 54% of the sample. White/Caucasian students were the most commonly reported ethnicity/race, with almost 59%. For analysis here, and in alignment with the demographic profile of the CS community, White and Asian students comprise our non-URM category, with Black/African–American, Hispanic/Latino, Native American/American Indian, multiracial, and other comprising URM (Beede et al., 2011; National Academies of Sciences Engineering and Medicine et al., 2018; Smith, Jagesic, Wyatt, & Ewing, 2018; Wiebe et al., 2018). Full sample demographics are reported in Table 2.

Standalone survey administration began in summer 2020 and included a virtual summer camp and remote classroom administration due to COVID-19. The virtual summer camp emphasized engineering topics for students in grades 3 to 5. Our survey served as a consent-only final activity the campers completed after a weeklong camp session. Remote survey administration through classroom teachers began in August 2020; none of the participating teachers were technology specialists, but rather 4th or 5th grade teachers who provided the parents with consent documents and followed up with consented and assented students. All of these students took the E-CSA instrument one time.

The classroom-based studies occurred in fall 2019 and February to March of 2020 (pre-pandemic) and participating students were expected to complete the E-CSA both before and after the intervention. The fall 2019 study involved block-based coding instruction across three pair programming conditions, assigned at the classroom level. The three conditions were traditional pair programming with one computer and related driver–navigator roles, two computers without roles, and two computers with roles (Vandenebrg, Rachmatullah, Lynch, Boyer, & Wiebe, 2021). This study included only 5th grade students and lasted four weeks. The spring 2020 study took place over five weeks and involved implementing and comparing four system-based features to encourage 4th and 5th grade students using traditional pair programming to transfer the driver–navigator roles appropriately and to talk to their partner more effectively.

3.3. Validation procedure

To answer Research Question 1, we conducted a validation of the developed instrument. The validation procedure used in this

Table 1
Elementary CS attitudes (E-CSA) instrument items.

Item number	Item wording	Construct
SE_1	I would like to use coding to make something new.	Self-efficacy
OE_1	If I learn coding, then I can improve things that people use everyday.	Outcome expectancy
SE_2	I am good at building code.	Self-efficacy
SE_3	I am good at fixing code.	Self-efficacy
OE_2	I am interested in what makes computer programs work.	Outcome expectancy
OE_3	Using code will be important in my future jobs.	Outcome expectancy
OE_4	I want to use coding to be more creative in my future jobs.	Outcome expectancy
OE_5	Knowing how to code computer programs will help me in math.	Outcome expectancy
OE_6	Knowing how to code computer programs will help me in engineering.	Outcome expectancy
OE_7	Knowing how to code computer programs will help me in science.	Outcome expectancy
SE_4	I believe I can be successful in coding.	Self-efficacy

Table 2
Participant self-reported demographics.

Individual-level variables	*N*	Percent
Age		
8	7	4.5
9	36	23.1
10	85	54.5
11	28	17.9
Gender		
Male	73	43.2
Female	92	54.4
No Response	4	2.4
Grade		
3rd	5	3
4th	52	30.8
5th	112	66.3
Race/Ethnicity		
White	100	59.2
Black/African–American	14	8.3
Hispanic/Latino	15	8.9
Asian	12	7.1
Native American/American Indian	3	1.8
Multiracial	12	7.1
Other	2	1.2

study was based on the Standards for Educational and Psychological Testing proposed by the American Educational Research Association, American Psychological Association, and National Council on Measurement in Education (American Educational Research Association, American Psychological Association, National Council on Measurement in Education, Joint Committee on Standards for Educational, & Psychological Testing (US), 2014). This standard suggests five significant points on which to validate an instrument: response processes, test content, internal structure, consequences of testing, criterion validity, or the relationship between the measured constructs and other theoretically related variables. Given that previous work addressed the first two points (Vandenberg et al., 2020), the current study focused on the three latter points.

A combination of the classical test theory and item response theory-Rasch approaches was used in this study to examine the psychometric model and internal structure of the E-CSA (cf. Rachmatullah, Akram et al., 2020; Rachmatullah, Wiebe et al., 2020). We started by evaluating the number of latent constructs (dimensionality) and the items' quality within the instrument. Two models were compared in this dimensionality analysis: one- and two-dimensional (factor, construct, and dimension used interchangeably throughout this paper) models. A one-dimensional model which groups all items in a single factor was used as the baseline. The two-dimensional model was based on our theoretical conceptualization of the instrument based on two factors: self-efficacy and outcome expectancy (Bandura, 1986; Eccles & Wigfield, 2002; Wigfield & Eccles, 2000). Adams and Wu's procedure was used to identify the best-fitting model which is the model that has the lowest values in final deviance across three criteria: Akaike Information Criterion (AIC), Akaike Information Criterion Corrected (AICc), and Bayesian Information Criterion (BIC). Mean-square (MNSQ) values were used to assess the item quality, with the assumption that a well-behaved high quality item has an MNSQ value ranging from 0.60 to 1.40 (Wright & Linacre, 1994). All of these analyses were run in ConQuest version 5.12.3 (Adams, Wu, Cloney, & Wilson, 2020).

Differential Item Functioning (DIF) was run on the instrument as part of our IRT methods to examine item bias. DIF analysis is used to evaluate whether a certain group member in the study has different probabilities of endorsing a certain item controlling for the overall score (Boone & Scantlebury, 2006; Boone, Staver, & Yale, 2013). An item exhibiting DIF indicates a bias toward a particular group (Boone & Scantlebury, 2006; Boone et al., 2013), such as with gender, males tend to agree more on items about sports than females. In other words, an individual item that is biased does not automatically warrant removal of the item, but an instrument that has DIF items may lead to interpretation issues with regards to the problematic demographic factor (Boone & Scantlebury, 2006; Boone et al., 2013). DIF analysis addresses what Messick (1995) called the generalizability aspect of construct validity. Gender (Male and Female) and majority group representation (URM vs. non-URM) are of considerable interest to the computer science education research and policy community (e.g. Belser et al., 2017; Beyer, 2014; Sax, Lehman et al., 2017; Sax, Zimmerman et al., 2017) and thus will be the focus of our DIF analysis. We used the cut-off value of < 0.64, as suggested by Boone et al. (2013), to evaluate the DIF of an item. An item that had a DIF contrast value more than the cutoff, it would be demonstrating unacceptable bias based on the demographic factor of interest and is typically removed.

After all the problematic items were removed, a CFA was then run to provide additional structural validity evidence. Informed by the results of the initial dimensionality analysis, we only ran CFA on the two-factor model and evaluated this model using the cut-off suggested by Hair, Black, Babin, and Anderson (2019). The acceptable model should have chi-square/df < 3, root mean

Table 3
Comparison between one- and two-dimensional models of E-CSA.

Model	χ^2	df	Final deviance	AIC	AICc	BIC	Num. of parameters	Num. of misfitting items
One-dimension	199.51	10	6521.66	6551.66	6549.95	6603.30	15	0
Two-dimension	209.08	9	6460.71	6494.71	6492.51	6553.23	17	0

Table 4
Item fit statistics for the two-dimensional E-CSA model.

Construct	Item code	Estimate	Weighted MNSQ	Unweighted MNSQ	DIF gender	DIF race	Alpha if item deleted
CS self-efficacy							
	SE_1	−0.369	1.02	0.94	0.21	0.27	0.794
	SE_2	0.306	0.89	0.85	0.09	0.15	0.716
	SE_3	0.541	1.05	1.05	0.06	0.07	0.754
	SE_4	−0.479	0.96	0.89	0.02	0.00	0.785
CS outcome expectancy							
	OE_1	−0.253	1.05	1.16	0.19	0.00	0.818
	OE_2	0.033	1.20	1.24	0.63	0.07	0.814
	OE_3	0.294	1.07	1.06	0.46	0.16	0.822
	OE_4	0.372	0.82	0.81	0.18	0.27	0.793
	OE_5	0.158	0.91	0.95	0.26	0.02	0.811
	OE_6	−0.490	1.23	1.18	0.46	**0.68**[a]	0.826
	OE_7	−0.114	1.09	1.13	0.49	0.19	0.828

[a]See Sections 5.1 and 4.1.2 for more about how to interpret results using this item.

square error of approximation (RMSEA) < .08, comparative fit index (CFI) > .95, and Tucker–Lewis index (TLI) > .95. CFA was performed in IBM SPSS Amos version 26 (Arbuckle, 2019).

Lastly, ensuring the internal consistency and accuracy of the item responses was done by evaluating the reliability values. Three different reliability values were computed using the CTT and IRT-Rasch methods, namely Cronbach's alpha, item separation reliability, and person reliability (person/a posteriori plausible value reliability). Linacre (2012) suggested that the two latter reliabilities can be evaluated in the same way as Cronbach's alpha. Thus we used the same cut-off value of > .70 to determine an acceptable reliability value (DeVellis, 2016).

3.3.1. Data analysis

Students' raw scores were converted using Rasch analysis to obtain scores in the ratio-interval form (logit). Each student had scores for both of the constructs, self-efficacy and outcome expectancy. We used these logit scores for the subsequent analyses. A Pearson correlation test was run to examine the relationship between CS attitudes – self-efficacy and outcome expectancy – and conceptual understanding of CS. An instrument, the E-CSCA (Elementary Computer Science Conceptual Assessment (Vandenebrg et al., 2021)) adopted from Rachmatullah, Wiebe et al. (2020) was used to measure students' conceptual understanding of CS. Two example items from the E-CSCA appear in the Appendix 5 and 6; these were based on the work of Rachmatullah, Wiebe et al. (2020) and recently validated in Vandenebrg et al. (2021). Furthermore, a two-way ANCOVA test was run to explore the interaction effect of gender and race/ethnicity on elementary students' CS self-efficacy and outcome expectancy, by controlling for test occasions (pre-posttest or standalone). Tests of simple slopes were run to decompose the interaction effect. The effect sizes were calculated using Cohen's d, with 0.20, 0.50 and 0.80 for small, medium and large effect sizes respectively (Lakens, 2013). All these analyses were performed using the **lm** package in RStudio (Team, 2020).

4. Results

4.1. Instrument validation

4.1.1. Multidimensional Rasch analysis and item fit-statistics

Multidimensional Rasch analysis was run to assess the best fitting model of E-CSA. Table 3 presents the results of the multidimensional Rasch analysis. We found that both one- and two-dimensional models of E-CSA did not have any misfitting items. However, the two-dimensional model had lower (i-e., better) values of the final deviance criteria (AIC, AICc, and BIC) than the one-dimensional model. A Chi-square test on the AIC showed a significant difference between one- and two-dimensional models (χ^2 = 56.95, p < .05), indicating that the two-dimensional model was the best model. We then used this two-dimensional model in our subsequent analysis.

Table 4 shows the fit statistics for all items in the two-dimensional model representing both constructs—CS self-efficacy and CS outcome expectancy. All the items had weighted and unweighted MNSQ values within the range of acceptable values, 0.60 −1.40, as suggested by Wright and Linacre (1994). These values demonstrated that the items were psychometrically sound and able to differentiate students based on the degree of their CS self-efficacy and outcome expectancy. Moreover, a Wright map (see Fig. 4) from the multidimensional analysis shows a reasonable distribution of students' CS self-efficacy and outcome expectancy responses, from strongly disagree (below Level 1) to strongly agree (above Level 4).

4.1.2. Differential item functioning — gender and race

We also ran DIF analyses for gender and race to address the generalizability aspect of construct validity. The results indicated that most of the items were free from gender and race-bias, suggesting that they behaved equally to all gender and race groups. We only detected one item with a DIF for race/ethnicity (URM/non-URM), OE_6, with a DIF contrast value of 0.83. We chose not to remove this item, as other psychometric indices indicated it was a good quality item. However, we suggest carefully interpreting the results using this item when conducting analyses comparing CS outcome expectancy by race/ethnicity. Table 4 presents the results of DIF gender and race/ethnicity analyses.

Table 5
Comparing CFA models with and without correlated residuals.

Indicator	Model 1	Model 2
df	42	38
χ^2/df(<3)	4.16	1.94
p-value	<.001	<.001
CFI (> .95)	.884	.969
TLI (>.95)	.818	.946
RMSEA (< .08)	.117	.064
$\Delta\ \chi^2$ (Δdf)	–	100.90
p-value for $\Delta\ \chi^2$	–	< .001

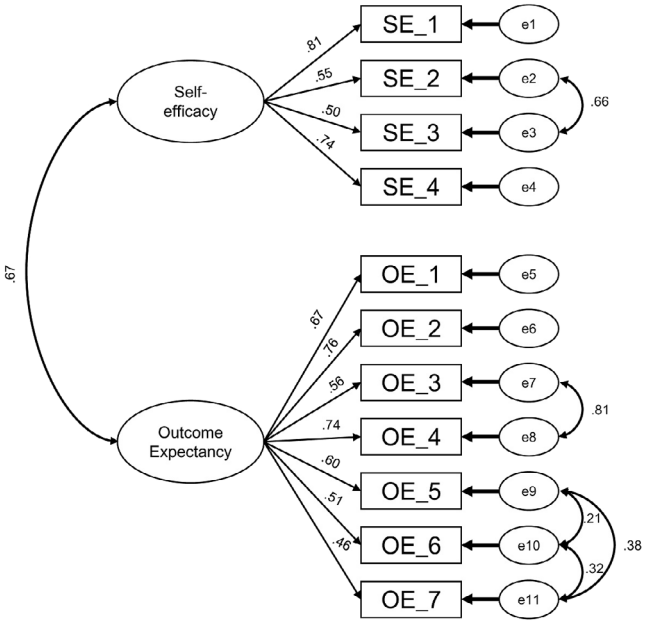

Fig. 1. Final CFA model for two-factor E-CSA with standardized loadings. Note: The figure above demonstrates that self-efficacy and outcome expectancy are two distinct factors (see Section 4.1.1), the individual items associated with the factors (see Table 1), and the correlated residuals indicated by the double headed arrows on the right (see Section 4.1.3 and Section 5.1).

4.1.3. CFA

The structural model of the E-CSA was then analyzed through CFA. All the original items were included in the CFA, as multidimensional Rasch and DIF analyses indicated no problematic items. We compared two models: the model without correlated residual errors (Model 1) and correlated residual errors (Model 2). For Model 2, the correlated residuals were determined based on the modification indices and the items' context (Hair et al., 2019). After evaluating all the fit statistics indicators, we found that Model 2 had significantly better fit statistics than Model 1. Table 5 shows all the fit statistics for these two models with Model 2 demonstrating lower chi-square and RMSEA and higher CFI and TLI, and therefore better values (see cut off values in Table 5 next to the indicators), and Fig. 1 visualizes the structure of the E-CSA two-factor model with correlated residual errors.

4.1.4. Reliability

Cronbach's alpha and plausible value (PV; aka person reliability) generated from the multidimensional Rasch analysis were used to assess the internal consistency of the E-CSA. The CS self-efficacy construct had Cronbach's alpha and PV reliability values of .812 and .843, respectively. For the CS outcome expectancy, the Cronbach's alpha and PV reliability values were .838 and .883, respectively. All of these values were above the acceptable value of .70 (DeVellis, 2016), indicating a stable instrument. Also, a separation reliability value was computed through multidimensional Rasch analysis, evaluating the reproducibility of the spread of the response levels. The separation reliability for the E-CSA was .960, indicating a good spread of item response.

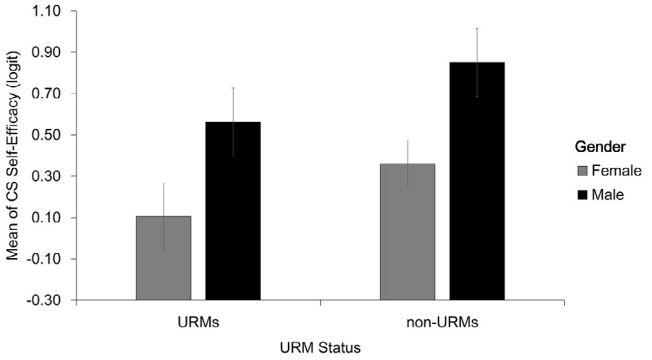

Fig. 2. Differences in CS self-efficacy based on gender and race.

4.2. Correlation between CS attitude and CS conceptual understanding

To address Research Question 2, Pearson correlation tests were run to examine the correlation between the two constructs in the E-CSA and students' conceptual understanding of CS, the E-CSCA. We found that CS self-efficacy had a significant positive correlation (r = .24, p = .002) with the CS conceptual understanding. In contrast, we did not find a significant correlation between CS outcome expectancy and CS conceptual understanding (r = .09, p = .278).

4.3. Interaction effect between gender and race on elementary CS attitude

To address Research Question 3, two-way ANCOVA tests were performed to examine the interaction effect of gender and race/ethnicity (based on URM vs. non-URM) on elementary students' CS self-efficacy and outcome expectancy. For CS self-efficacy, we found that the interaction effect between gender and race/ethnicity was not significant after controlling for test occasion (pre-posttest or standalone; t = 0.03, p = .920). We then removed the interaction effect from the model, and ran another model in which we found that gender had a significant fixed effect on elementary students' CS self-efficacy with a small effect size (t = 3.15, p = .002, d = .11). Decomposing this result, male students (M = 0.79, SD = 0.69) had higher CS self-efficacy than female students (M = 0.32, SD = 0.93). In contrast, there was a non-significant fixed effect of race/ethnicity on CS self-efficacy (t = 0.27, p = .11, d = 0.22) indicating non-URM students (M = 0.55, SD = 1.21) did not differ from URM students (M = 0.36, SD = 0.90). The results are visualized in Fig. 2.

Similar to the findings in CS self-efficacy, the interaction effect of gender and race/ethnicity was not significant on the CS outcome expectancy (t = 0.17, p = .577). After we removed this interaction effect from the model, we again found a significant fixed effect of gender on CS outcome expectancy (t = 3.05, p = .002, d = .04), where male students (M = 0.69, SD = 1.02) had higher scores than female students (M = 0.27, SD = 0.93). As with the findings for CS self-efficacy, we also did not find a significant fixed effect of race/ethnicity on CS outcome expectancy (t = 0.16, p = .25, d = .33). This indicated that non-URM students (M = 0.47, SD = 1.02) did not differ from URM students (M = 0.37, SD = 0.92) in CS outcome expectancy. Fig. 3 presents the results.

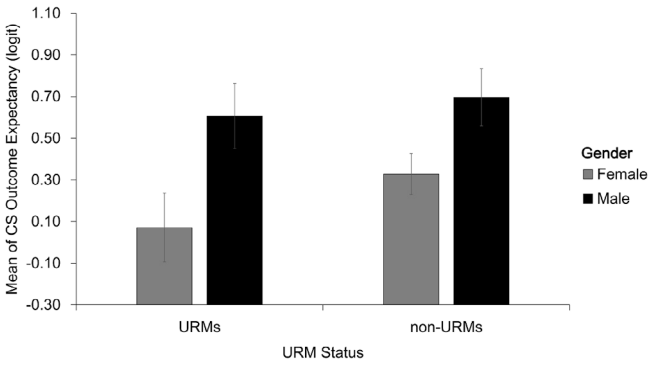

Fig. 3. Differences in CS outcome expectancy based on gender and race.

5. Discussion

In order to address gaps in CS participation, from elementary classrooms to university major enrollment, it is important to explore students' attitudes (self-efficacy and outcome expectancy) toward CS, and to what extent differences in attitude appear by race/ethnicity and gender. As such, we set out to examine these differences through a brief bias-free instrument we developed and validated, and appropriate for upper elementary student use. We discuss our findings by research question.

5.1. Research Question 1: With regard to the validation of the instrument, what model best represents the dimensionality and internal structure of the E-CSA?

In this study, we achieved content validity by relying on prior scholarly work on self-efficacy (Bandura et al., 1999) and outcome expectancy (Eccles & Wigfield, 2002; Wigfield & Eccles, 2000) and by utilizing and modifying a previously validated instrument (S-STEM Unfried, Faber, Stanhope, & Wiebe, 2015). Moreover, prior work of ours made use of the rigorous process of cognitively interviewing a diverse array of upper elementary students on their understanding of the terminology used in the instrument items (Vandenberg et al., 2020). Others (Padilla & Benítez, 2014; Willson & Miller, 2014) have utilized this approach to ensure content validity based on response processes.

Regarding consistencies in test responses, as typically evaluated through reliability values, we utilized Cronbach's alpha and PV reliability values generated through CTT and IRT methods. Our results indicate a stable instrument in which participants consistently responded to the items within each factor in a relatively similar fashion.

Additionally, we explored the instrument and individual item quality through confirmatory factor analysis (CFA) and multidimensional Rasch modeling. Having established an *a priori* hypothesis about the latent factors and variables, owing in large part to basing our work on a previously validated and theoretically-sound instrument, we were confident in beginning our classical test theory work with a CFA (e.g. Thompson, 2004). Our CFA and multidimensional Rasch modeling resulted in a two-factor model as best fit, aligning with our a priori expectations based on the theoretical framework upon which the instrument is based. These assumptions were supported, as the two-factor model with correlated residuals had significantly better fit statistics. This type of model fitting is used when theoretically and meaningfully justified (Brown, 2003, 2015; Cole, Ciesla, & Steiger, 2007), such as when covariance occurs due to content overlap or item phrasing. In our case, we identified sets of items whose content/wording and sequential proximity may have influenced how students responded to them. In particular, SE_2 and SE_3, "I am good at building code" and "I am good at fixing code", respectively, appear in sequence and have extremely similar wording. Additionally, OE_3 and OE_4, "Using code will be important in my future jobs" and "I want to use coding to be more creative in my future jobs", respectively, also appear in sequence and both reference "future jobs". Lastly, we permitted the residuals of the following to correlate: "Knowing how to code computer programs will help me in ——" math (OE_5), engineering (OE_6), and science (OE_7). By allowing these terms to correlate, our fit indices improved to acceptable levels.

Lastly, regarding generalizability, we completed DIF analysis to explore the fairness of the instrument across varied socio-demographic subgroups of students. Our results indicate that the instrument is largely free, psychometrically, from gender and race bias, with one item (OE_6) demonstrating marginal DIF by race/ethnicity. This was near the threshold for removal (Boone et al., 2013); we opted to retain the item, but users of the instrument need to be aware. This item, "Knowing how to code computer programs will help me in engineering", was one where we found marked qualitative differences in students' responses during the cognitive interview process. Based on our prior work (Vandenberg et al., 2020), students in rural, under-served, low SES, and largely Black/African–American and Hispanic/Latino schools struggled to provide a robust definition for engineering. Therefore, the scores computed from E-CSA's outcome expectancy scale should be carefully interpreted when comparing groups based on race/ethnicity on this scale. We believe that this problematic item highlights the need for more substantive exposure to engineering education and experiences at the elementary level for all children.

5.1.1. Research Question 2: What is the relationship between elementary students' CS attitudes and their CS conceptual understanding?

In this study, we treated the students' scores on a measure of conceptual CS understanding as being theoretically related to CS self-efficacy and outcome expectancy measured via E-CSA. Our results indicate that of the two factors that comprise the E-CSA, only self-efficacy was significantly positively correlated with conceptual understanding. Self-efficacy has long been considered a predictor of student outcomes (Bandura, 1986; Brosnan, 1998; Lishinski et al., 2016); these empirical findings align with our own. Research indicates that there is positive impact of (CS-related) experience on self-efficacy (Bandura, 1986; Hinckle et al., 2020). Further, by improving CS self-efficacy, we can expect to see improvements in CS conceptual understanding. Although outcome expectancy was not correlated with student scores on the E-CSA, we believe it is still a valuable measure, as it compliments self-efficacy in providing a more complete motivational model of the student with regards to CS (Eccles & Wigfield, 2002).

5.1.2. Research Question 3: What is the influence of race/ethnicity and gender on students' responses on the E-CSA instrument?

We found that gender had a statistically significant effect on CS Attitudes, with males having higher self-efficacy and outcome expectancy than females. This difference has profound implications for both proximal and distal interests and performance and it is not a new issue. Empirical research from the 1990s indicated that males self-report higher confidence for, more liking of, and lower anxiety with computers (Charlton, 1999; Colley, Gale, & Harris, 1994). Newer research largely mirrors earlier findings (Beyer, 2014; Wilcox & Lionelle, 2018), although there may be reason to hope as these findings likely indicate a path forward for girls. In particular, Schmidt (2011) found that females' lower interest in technology leads to reduced experience and

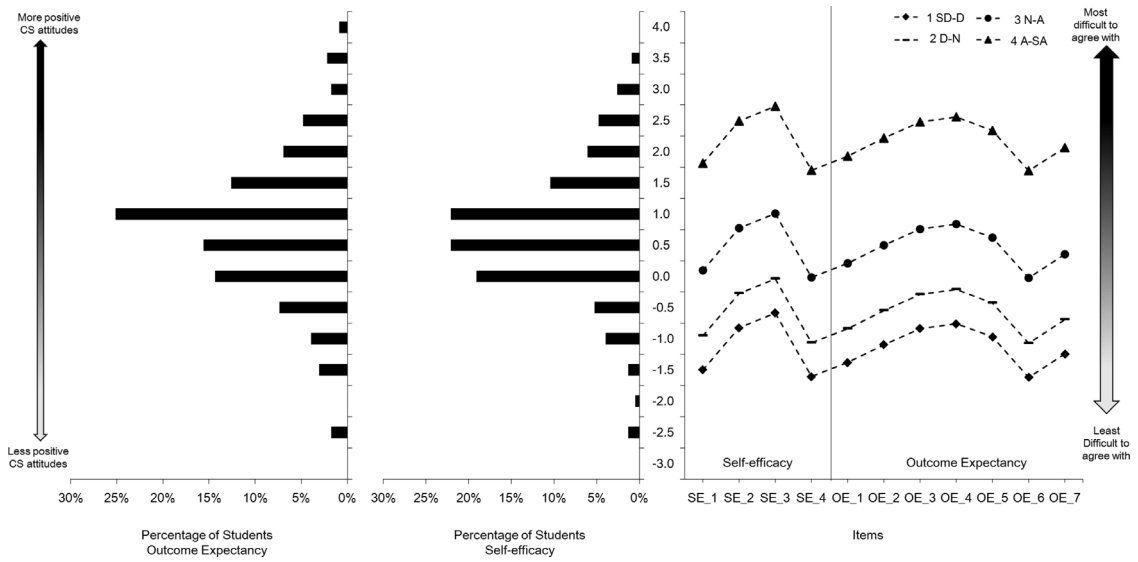

Fig. 4. The Wright map for The E-CSA showing agreement difficulty for each item and students' attitudinal spectrum. Note: The agreement difficulties for each item's scales on the right are represented with Thurstonian thresholds, which refer to a specific location where a student has a 50% probability of choosing a given scale or higher. Students' CS attitudes are represented with histograms on the left. When a student appears to be precisely aligned with a Thurstonian threshold, this means that the student has an equal probability of selecting the scale or option above or below the threshold. SD = Strongly Disagree, D = Disagree, N = Neutral, A = Agree, SA = Strongly Agree).

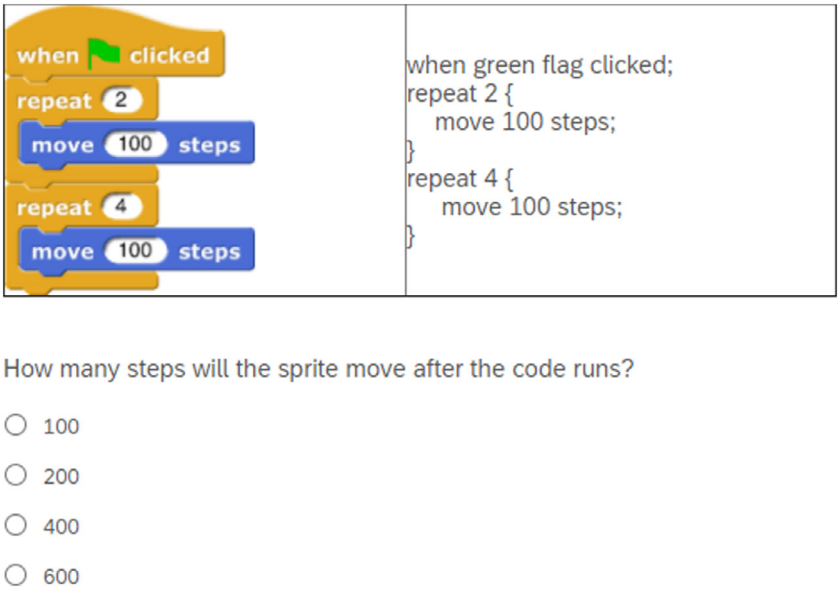

Fig. 5. Example item from E-CSCA.

knowledge. Contrast that with Wilcox and Lionelle (2018) who note that female students outperform their male peers when they have similar levels of prior computing experience. Providing girls and young women with consistent and quality computing experiences is essential. This needs to be addressed through concerted efforts at even younger grades than we tested here in order to try to prevent the development of these deleterious gender-based attitudes. In addition, it is noteworthy that most previous studies investigating gender differences in elementary or upper education level students' attitudes toward CS did not examine whether the instrument used in those studies were free from bias (cf. Kong et al., 2018; Mason & Rich, 2020). Thus, the results may not be valid with regards to research questions centering on gender. We believe our findings on gender differences in elementary CS attitudes toward CS have rigorously addressed this potential problem with item bias, as DIF analysis indicated that our instrument items were free of gender bias.

It is also important to note that there was a nonsignificant effect of race/ethnicity on CS self-efficacy and outcome expectancy. Based on our findings, URM and non-URM students did not differ in these constructs. This is meaningful as other research indicates that URMs often indicate lower interest in CS and generally find CS to be an unwelcoming place (Margolis et al., 2017; Scott et al., 2017). That we did not find a statistically significant difference is intriguing. It could be that these young students have not yet encountered negative racial and ethnic stereotypes that might influence their perceptions of themselves. Most prior research on racial and ethnic stereotypes have occurred with older students (e.g., Johnson, 2011; Margolis et al., 2017; Scott et al., 2017); however, a recent study found that young children, ages 3 to 8, did not use racial/ethnic information to make decisions about

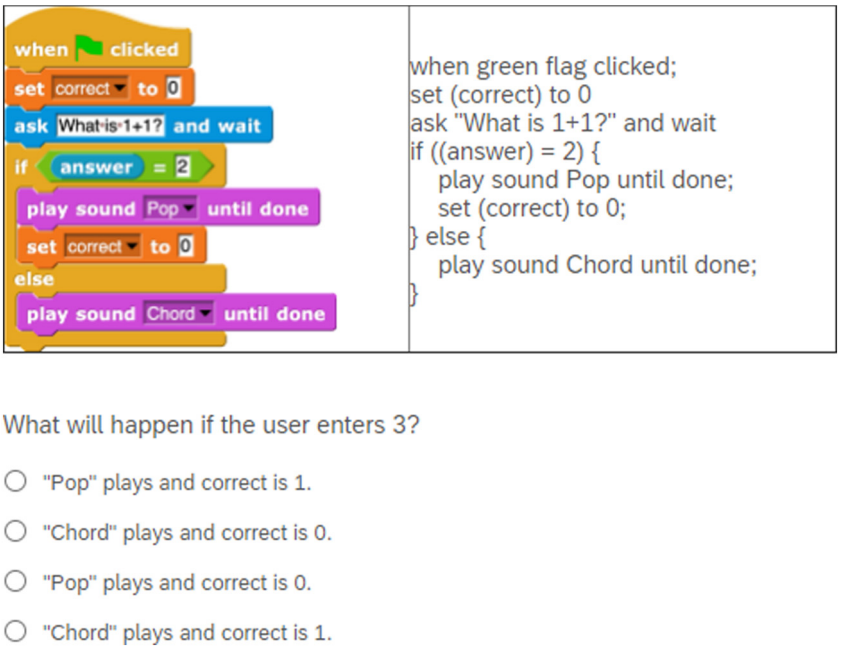

Fig. 6. Example item from E-CSCA.

who ought to perform certain STEM-based jobs (Mulvey & Irvin, 2018).

6. Limitations

We acknowledge the following limitations of this study and suggest them for future work as part of both the instrument development process and examining gender and race/ethnicity in CS education. First, we had a relatively small sample size (N = 169) of students who completed the E-CSA instrument. To compensate for this, we used robust, psychometrically sound techniques for this smaller sample size. However, the students who did participate were largely white (59.2%) and thus limited our ability to explore the relationship of race/ethnicity to both attitudinal and learning factors. Future work would benefit from a more diverse racial and ethnic sample. It is worth noting that grouping by URM and non-URM is not the only approach that can be used to examine effects of race and ethnicity. While it increases sub-sample size (and related statistical power) to group multiple demographic categories, it can also mask important patterns happening at a finer-grained level. Relatedly, despite purposeful sampling across diverse school populations and contexts, all results are from a single state in the United States. Future work would benefit from more widespread national and international data collection and with populations with various levels of CS-related experiences. Additionally, we did not account for teacher- or school-level differences; future work with a more substantive sample size might consider conducting multilevel modeling to explore this further (Lee, 2000). Also, survey item order was set, perhaps contributing to the nonrandom errors in the model. Future administrations should consider randomizing the items to test for and reduce this effect. Finally, we recognize that we have not fully captured upper elementary students' attitudes toward and knowledge of computer science, so further additions and refinement are suggested.

7. Conclusion

This study examined gender and race differences in elementary students' attitudes toward CS. To that end, we developed and validated a survey called Elementary Computer Science Attitudes (E-CSA) which consisted of the constructs of CS self-efficacy and outcome expectancy, through a combination of classical test theory (CTT) and item response theory (IRT) Rasch. The E-CSA was found to be, psychometrically, a gender and race bias-free instrument. We found no significant interaction effect between gender and race in the two constructs of CS Attitudes. We also did not see a significant difference based on race. However, a significant difference was found in both CS attitudes constructs based on gender, whereby male students had higher CS attitudes than female students.

Prior work has established the link between students' beliefs, such as their self-efficacy for a content area, and their performance in that area (Brosnan, 1998; Lishinski et al., 2016). Having an instrument that assesses students' attitudes toward computer science that is based on theoretically-derived constructs, self-efficacy and outcome expectancy, could prove indispensable to researchers and practitioners alike. To this end, we developed and rigorously validated a brief instrument appropriate for use with upper elementary students.

We believe that use of the instrument can inform classroom-based interventions, the development of curricular materials, and reinforce findings from other cross-sectional CS studies. In particular, we believe that our findings support the need for early and consistent CS interventions with girls (Happe, Buhnova, Koziolek, & Wagner, 2020; Hur, Andrzejewski, & Marghitu, 2017) so as to support their positive attitudes toward CS. Moreover, as the instrument was validated with upper elementary students, we support the use of it alongside other analyses with the same aged population.

In addition to addressing the limitations noted above, future research could explore how CS attitudes correlate with non-STEM subject areas. Prior work of ours indicated that some students understood the CS concept of debugging to be much like editing and revising a paper in a writing class. Additionally, we are interested in how remote learning and the increased use of technology may have implications for students' interest in CS. And lastly, future work could explore how students' talk about CS and coding may reflect their beliefs and overall interests.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

This material is based upon work supported by the National Science Foundation, USA under Grant No. DRL-1721160. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

Appendix A

See Fig. 4.

Appendix B

See Figs. 5 and 6.

References

Adams, R., & Wu, M. (2010). Notes and tutorial conquest: Multidimensional model.

Adams, R., Wu, M., Cloney, D., & Wilson, M. (2020). ACER ConQuest: generalised item response modelling software [Computer software] Version 4. Australian council for educational research, Camberwell.

Aivaloglou, E., & Hermans, F. (2019). Early programming education and career orientation: the effects of gender, self-efficacy, motivation and stereotypes. In *Proceedings of the 50th ACM technical symposium on computer science education* (pp. 679–685).

Aladé, F., Lauricella, A., Kumar, Y., & Wartella, E. (2020). Who's modeling STEM for kids? A character analysis of children's STEM-focused television in the US. *Journal of Children and Media*, 1–20.

American Educational Research Association, American Psychological Association, National Council on Measurement in Education, Joint Committee on Standards for Educational, & Psychological Testing (US) (2014). *Standards for educational and psychological testing*. Amer Educational Research Assn.

Arbuckle, J. (2019). Amos (Version 26.0)[Computer Program] Chicago. IL, USA: IBM SPSS.

Bandura, A. (1986). *Social foundations of thought and action* (pp. 23–28). Englewood Cliffs, NJ.

Bandura, A., Freeman, W., & Lightsey, R. (1999). *Self-efficacy: The exercise of control*. Springer.

Beede, D. N., Julian, T. A., Khan, B., Lehrman, R., McKittrick, G., Langdon, D., et al. (2011). Education supports racial and ethnic equality in STEM. Economics and Statistics Administration Issue Brief 05–11.

Belser, C. T., Prescod, D. J., Daire, A. P., Dagley, M. A., & Young, C. Y. (2017). Predicting undergraduate student retention in STEM majors based on career development factors. *The Career Development Quarterly*, *65*(1), 88–93.

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.

Beyer, S., Rynes, K., Perrault, J., Hay, K., & Haller, S. (2003). Gender differences in computer science students. In *Proceedings of the 34th SIGCSE technical symposium on computer science education* (pp. 49–53).

Boone, W. J., & Scantlebury, K. (2006). The role of rasch analysis when conducting science education research utilizing multiple-choice tests. *Science Education*, *90*(2), 253–269.

Boone, W. J., Staver, J. R., & Yale, M. S. (2013). *Rasch analysis in the human sciences*. Springer.

Brosnan, M. J. (1998). The impact of computer anxiety and self-efficacy upon performance. *Journal of Computer Assisted Learning*, *14*(3), 223–234.

Brown, T. A. (2003). Confirmatory factor analysis of the Penn State Worry Questionnaire: Multiple factors or method effects?. *Behaviour Research and Therapy*, *41*(12), 1411–1426.

Brown, T. A. (2015). *Confirmatory Factor Analysis for Applied Research*. Guilford publications.

Buffum, P. S., Frankosky, M., Boyer, K. E., Wiebe, E., Mott, B., & Lester, J. (2015). Leveraging collaboration to improve gender equity in a game-based learning environment for middle school computer science. In *2015 research in equity and sustained participation in engineering, computing, and technology (RESPECT)* (pp. 1–8). IEEE.

Charlton, J. P. (1999). Biological sex, sex-role identity, and the spectrum of computing orientations: A re-appraisal at the end of the 90s. *Journal of Educational Computing Research*, *21*(4), 393–412.

Chen, W.-H., Lenderking, W., Jin, Y., Wyrwich, K. W., Gelhorn, H., & Revicki, D. A. (2014). Is rasch model analysis applicable in small sample size pilot studies for assessing item characteristics? An example using PROMIS pain behavior item bank data. *Quality of Life Research*, *23*(2), 485–493.

Cheryan, S., Master, A., & Meltzoff, A. N. (2015). Cultural stereotypes as gatekeepers: Increasing girls' interest in computer science and engineering by diversifying stereotypes. *Frontiers in Psychology*, *6*, 49.

Cheryan, S., Plaut, V. C., Davies, P. G., & Steele, C. M. (2009). Ambient belonging: how stereotypical cues impact gender participation in computer science. *Journal of Personality and Social Psychology*, *97*(6), 1045.

Cole, D. A., Ciesla, J. A., & Steiger, J. H. (2007). The insidious effects of failing to include design-driven correlated residuals in latent-variable covariance structure analysis. *Psychological Methods*, *12*(4), 381.

Colley, A. M., Gale, M. T., & Harris, T. A. (1994). Effects of gender role identity and experience on computer attitude components. *Journal of Educational Computing Research*, *10*(2), 129–137.

DeVellis, R. (2016). *Scale development: theory and applications* (4th ed.). Thousand Oaks (CA): Sage Publications Inc.

Eccles, J. S., & Wigfield, A. (2002). Motivational beliefs, values, and goals. *Annual Review of Psychology*, *53*(1), 109–132.

Friday Institute for Educational Innovation (2012). Student attitudes toward STEM survey—Middle and high school students.

Google Inc., & Gallup Inc. (2016). Diversity gaps in computer science: Exploring the underrepresentation of girls, blacks and hispanics.

Hair, J. E., Jr., Black, W., Babin, B., & Anderson, R. (2019). *Multivariate data analysis* (8th ed.). United Kingdom: Cengage Learning EMEA.

Happe, L., Buhnova, B., Koziolek, A., & Wagner, I. (2020). Effective measures to foster girls' interest in secondary computer science education. *Education and Information Technologies*, 1–19.

Hinckle, M., Rachmatullah, A., Mott, B., Boyer, K. E., Lester, J., & Wiebe, E. (2020). The relationship of gender, experiential, and psychological factors to achievement in computer science. In *Proceedings of the 2020 ACM conference on innovation and technology in computer science education* (pp. 225–231).

Hur, J. W., Andrzejewski, C. E., & Marghitu, D. (2017). Girls and computer science: experiences, perceptions, and career aspirations. *Computer Science Education*, *27*(2), 100–120.

James DiSalvo, B., Yardi, S., Guzdial, M., McKlin, T., Meadows, C., Perry, K., et al. (2011). African American men constructing computing identity. In *Proceedings of the SIGCHI conference on human factors in computing systems* (pp. 2967–2970).

Johnson, D. R. (2011). Examining sense of belonging and campus racial diversity experiences among women of color in STEM living-learning programs. *Journal of Women and Minorities in Science and Engineering*, *17*(3).

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.

Kukul, V., Gökçearslan, Ş., & Günbatar, M. S. (2017). Computer programming self-efficacy scale (CPSES) for secondary school students: Development, validation and reliability. *Eğitim Teknolojisi Kuram ve Uygulama*, *7*(1), 158–179.

Lakens, D. (2013). Calculating and reporting effect sizes to facilitate cumulative science: a practical primer for t-tests and ANOVAs. *Frontiers in psychology*, *4*, 863.

Leath, S., & Chavous, T. (2018). Black women's experiences of campus racial climate and stigma at predominantly white institutions: Insights from a comparative and within-group approach for STEM and non-STEM majors. *The Journal of Negro Education*, *87*(2), 125–139.

Lee, V. E. (2000). Using hierarchical linear modeling to study social contexts: The case of school effects. *Educational Psychologist*, *35*(2), 125–141.

Lent, R. W., Lopez, A. M., Jr., Lopez, F. G., & Sheu, H.-B. (2008). Social cognitive career theory and the prediction of interests and choice goals in the computing disciplines. *Journal of Vocational Behavior*, *73*(1), 52–62.

Linacre, J. (1994). Sample size and item calibration stability. *Rasch Measurement Transactions*, *7*, 328.

Linacre, J. (2012). A user's guide to Winstep. Ministep Rasch-Model Computer Programs: Program Manual by John M. Linacre. Chicago. Winsteps. com.

Lishinski, A., Yadav, A., Good, J., & Enbody, R. (2016). Learning to program: Gender differences and interactive effects of students' motivation, goals, and self-efficacy on performance. In *Proceedings of the 2016 ACM conference on international computing education research* (pp. 211–220).

Margolis, J., Estrella, R., Goode, J., Holme, J. J., & Nao, K. (2017). *Stuck in the shallow end: Education, race, and computing*. MIT press.

Mason, S. L., & Rich, P. J. (2020). Development and analysis of the elementary student coding attitudes survey. *Computers & Education*, Article 103898.

Master, A., Cheryan, S., Moscatelli, A., & Meltzoff, A. N. (2017). Programming experience promotes higher STEM motivation among first-grade girls. *Journal of Experimental Child Psychology*, *160*, 92–106.

Master, A., & Meltzoff, A. N. (2020). Cultural stereotypes and sense of belonging contribute to gender gaps in STEM. *International Journal of Gender, Science and Technology*, *12*(1), 152–198.

McDowell, C., Werner, L., Bullock, H. E., & Fernald, J. (2006). Pair programming improves student retention, confidence, and program quality. *Communications of the ACM*, *49*(8), 90–95.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning.. *American Psychologist*, *50*(9), 741.

Muenks, K., Wigfield, A., & Eccles, J. S. (2018). I can do this! the development and calibration of children's expectations for success and competence beliefs. *Developmental Review*, *48*, 24–39.

Mulvey, K. L., & Irvin, M. J. (2018). Judgments and reasoning about exclusion from counter-stereotypic STEM career choices in early childhood. *Early Childhood Research Quarterly*, *44*, 220–230.

National Academies of Sciences Engineering and Medicine, et al. (2018). *Assessing and responding to the growth of computer science undergraduate enrollments*. National Academies Press.

Padilla, J.-L., & Benítez, I. (2014). Validity evidence based on response processes. *Psicothema*, *26*(1), 136–144.

Pajares, F. (1996). Self-efficacy beliefs and mathematical problem-solving of gifted students. *Contemporary Educational Psychology*, *21*(4), 325–344.

Plante, I., De la Sablonnière, R., Aronson, J. M., & Théorêt, M. (2013). Gender stereotype endorsement and achievement-related outcomes: The role of competence beliefs and task values. *Contemporary Educational Psychology*, *38*(3), 225–235.

Rachmatullah, A., Akram, B., Boulden, D., Mott, B., Boyer, K., Lester, J., et al. (2020). Development and validation of the middle grades computer science concept inventory (MG-CSCI) assessment. *EURASIA Journal of Mathematics, Science and Technology Education*, *16*(5), em1841.

Rachmatullah, A., Wiebe, E., Boulden, D., Mott, B., Boyer, K., & Lester, J. (2020). Development and validation of the Computer Science Attitudes Scale for middle school students (MG-CS attitudes). *Computers in Human Behavior Reports*, *2*, Article 100018.

Sax, L. J., Lehman, K. J., Jacobs, J. A., Kanny, M. A., Lim, G., Monje-Paulson, L., et al. (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.

Sax, L. J., Zimmerman, H. B., Blaney, J. M., Toven-Lindsey, B., & Lehman, K. (2017). Diversifying undergraduate computer science: The role of department chairs in promoting gender and racial diversity. *Journal of Women and Minorities in Science and Engineering*, *23*(2).

Schmidt, F. L. (2011). A theory of sex differences in technical aptitude and some supporting evidence. *Perspectives on Psychological Science*, *6*(6), 560–573.

Scott, A., Martin, A., McAlear, F., & Koshy, S. (2017). Broadening participation in computing: examining experiences of girls of color. In *Proceedings of the 2017 ACM conference on innovation and technology in computer science education* (pp. 252–256).

Smith, K., Jagesic, S., Wyatt, J., & Ewing, M. (2018). AP® STEM participation and postsecondary STEM outcomes: Focus on underrepresented minority, first-generation, and female students. *College Board*.

Team, R. S. (2020). R Studio: Integrated Development for R, 1.2. 5033. R Studio, PCB.

Thompson, B. (2004). *Exploratory and confirmatory factor analysis*. American Psychological Association.

Tsan, J., Boyer, K. E., & Lynch, C. F. (2016). How early does the CS gender gap emerge? A study of collaborative problem solving in 5th grade computer science. In *Proceedings of the 47th ACM technical symposium on computing science education* (pp. 388–393).

Unfried, A., Faber, M., Stanhope, D. S., & Wiebe, E. (2015). The development and validation of a measure of student attitudes toward science, technology, engineering, and math (S-STEM). *Journal of Psychoeducational Assessment*, *33*(7), 622–639.

Vandenberg, J., Tsan, J., Boulden, D., Zakaria, Z., Lynch, C., Boyer, K. E., et al. (2020). Elementary students' understanding of CS terms. *ACM Transactions on Computing Education (TOCE)*, *20*(3), 1–19.

Vandenebrg, J., Rachmatullah, A., Lynch, C., Boyer, K. E., & Wiebe, E. (2021). The relationship of CS attitudes, perceptions of collaboration, and pair programming strategies on upper elementary students' CS learning. In *Proceedings of the 2021 ACM conference on innovation and technology in computer science education*.

Werner, L. L., Hanks, B., & McDowell, C. (2004). Pair-programming helps female computer science students. *Journal on Educational Resources in Computing (JERIC)*, *4*(1), 4–es.

Wiebe, E., Unfried, A., & Faber, M. (2018). The relationship of STEM attitudes and career interest. *EURASIA Journal of Mathematics, Science and Technology Education*, *14*(10).

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

Wilcox, C., & Lionelle, A. (2018). Quantifying the benefits of prior programming experience in an introductory computer science course. In *Proceedings of the 49th ACM technical symposium on computer science education* (pp. 80–85).

Willson, S., & Miller, K. (2014). Data collection. *Cognitive interviewing methodology*. John Wiley & Sons.

Wright, B., & Linacre, J. (1994). Reasonable mean-square fit values. Rasch Measurement Transactions [On-line]. Retrieved November 30, 2009.

Yoo, M. H., Choi, J. J., Park, M. S., Chae, S. J., Kim, B. R., Son, M. H., et al. (2017). The effects of 'silver care expert' future career-related STEAM program for the elementary students' future time perspective, career awareness and scientific attitudes. *Journal of Science Education*, *41*(1), 111–134.

Zeldin, A. L., Britner, S. L., & Pajares, F. (2008). A comparative study of the self-efficacy beliefs of successful men and women in mathematics, science, and technology careers. *Journal of Research in Science Teaching: The Official Journal of the National Association for Research in Science Teaching*, *45*(9), 1036–1058.