Computers in Human Behavior Reports 2 (2020) 100018

Contents lists available at ScienceDirect

Computers in Human Behavior Reports

journal homepage: www.journals.elsevier.com/computers-in-human-behavior-reports

Development and validation of the Computer Science Attitudes Scale for middle school students (MG-CS attitudes)

Arif Rachmatullah [a,*], Eric Wiebe [a], Danielle Boulden [a], Bradford Mott [a], Kristy Boyer [b], James Lester [a]

[a] North Carolina State University, Raleigh, NC, USA
[b] University of Florida, Gainesville, FL, USA

ARTICLE INFO

Keywords:
computer Science attitudes
Invariance test
Scale validation
Secondary education

ABSTRACT

Primary and secondary students' exposure to computer science-related activities in school has been steadily increasing, heightening the need for valid measures regarding impact of these activities on students. This study reports on the development and validation process of an instrument to measure students' affective state as it relates to computer science in an academic setting. The self-report instrument, Computer Science Attitudes Scale for middle school students (MG-CS Attitudes), was developed based upon expectancy-value theory, which assumes two attitudinal constructs: self-efficacy and outcome expectancy. A set of ten initial items was administered to 663 middle-grade students from sixth to eighth grade (11–13 years of age). A combination of classical test theory and item response theory approaches were used to evaluate and validate the instrument using well-established construct validity frameworks to guide the process, leading to nine final items. The multi-stage validation process has resulted in a robust, well-functioning instrument, which can be used by researchers and evaluators to study CS-related educational interventions.

1. Introduction

1.1. Background

The past decade has seen unprecedented growth in industries and services that depend on computational capabilities provided by computer scientists and other STEM professionals trained in the use of computational and data-intensive tools, techniques, and theory (Stanton et al., 2017). However, this growth has not benefited all who are capable or aspire to do this computationally-intensive work. Large segments of the United States (U.S.) population, including women and people from historically marginalized groups such as African-Americans and Hispanic/Latinx, are not adequately represented in computationally-intensive STEM professions and in the higher education degree programs that train them (Google and Gallup, 2016; Levitan, 2018).

Middle grades (ages 11–13) have been identified as a critical age for engaging students, especially females and students from historically marginalized groups, in computational thinking (CT) and computer science (CS; Denner et al., 2012; Grover et al., 2016). Research has explored how we can increase student retention (Basawapatna et al., 2010) and improve learning for CT and CS (Rachmatullah et al., 2020; Wilkerson-Jerde et al., 2015) using approaches ranging from game-based learning to robotics and music (e.g., Sharek & Wiebe, 2014; Edwards, 2011).

Exposure to computational activities in school can increase students' interest and ability in computational thinking and computer science, which in turn can shape their motivation to engage in computing activities in the future and perhaps even to consider careers in computer science. Researchers and curriculum developers have sought to design curricula and learning experiences that positively impact students' computer science (CS) attitudes, especially during these crucial, formative middle-grade years (Lewis, 2010). Thus, attitudinal assessment is an essential component of curriculum development as it helps to verify the effectiveness of new curriculum interventions (Porter, 2006).

To date, there are some published validation studies on the development of CS-attitudes scales (e.g., Korkmaz et al., 2017; Tsai et al., 2019). However, much of this work is not guided by foundational psychological theory nor does it utilize robust psychometric validation

* Corresponding author.
E-mail addresses: arachma@ncsu.edu (A. Rachmatullah), wiebe@ncsu.edu (E. Wiebe), dmboulde@ncsu.edu (D. Boulden), bwmott@ncsu.edu (B. Mott), kristy@learndialogue.org (K. Boyer), lester@ncsu.edu (J. Lester).

https://doi.org/10.1016/j.chbr.2020.100018
Received 26 November 2019; Received in revised form 11 May 2020; Accepted 26 May 2020
Available online xxxx

methods. The published instruments focused more on measuring attitudes towards the CS constructs, such as algorithm, conditional, and control (e.g., Tsai et al., 2019). We argue that these specific constructs, as well as the vernacular associated with them, are too complex for middle school students to respond to in a meaningful manner. In addition, a construct-level approach misses the goal of capturing a holistic affective response to CS rather than to specific topical areas within the field. Finally, construct level assessment works at a granularity that greatly lengthens the instrument, creating the risk of subject fatigue and non-compliance that far outweighs any additional information that might be gained (Groves et al., 2009; Tourangeau et al., 2000).

In response, we followed previous studies on STEM attitudes, such as the studies conducted by Else-Quest et al. (2013), Summers and Abd-El-Khalick (2018), and Shin et al. (2018), where they define attitudes as psychological constructs that consist of, but not limited to, self-efficacy and outcome expectancy. In this current study, we focused on these two constructs suggested by Osborne et al. (2003) as considered core components of education-related attitudes. They also add that these constructs are easily recognized by students in self-report questionnaires. Building off of these initial works, we report on the development and validation process of an English language instrument for measuring attitudinal states related to CS educational experiences for middle-grade students (MG-CS Attitudes) that utilizes a theoretically guided validation process.

1.2. Theoretical framing

The goal in developing this instrument is to give researchers and evaluators a short, self-report instrument that would provide insight as to the impact of educational interventions intended to improve students' interest in engaging in CS-related educational activities. Our study attempted to address this need by utilizing a well-accepted psychological theory – expectancy-value theory (Eccles and Wigfield, 2002; Wigfield and Eccles, 2000). A modern social psychology model, expectancy-value theory, takes the fundamental notions of self-determination theory (Deci and Ryan, 1985) and places them within a goal-directed environment, such as academic and career pathways. Expectancy-value theory also leverages Bandura's (1986) conceptualization of self-efficacy, *the belief in one's ability to complete tasks or influence events that have an impact on one's life*. Self-efficacy has two broad facets: general self-efficacy to face challenges and tasks and self-efficacy specific to a particular task domain (Chen et al., 2001). Drawing on the latter, expectancy-value theory helps frame self-efficacy and outcome expectancy in terms of prospects of success in a particular academic domain and the value of this academic subject area in relation to future goals. Expectancy-value theory expands on outcome expectancy in academic work by positing that achievement-related performance and future academic or career choices are most directly influenced by the individual's expectations of academic success and the subjective assessment of the inherent value of these academic tasks (Eccles and Wigfield, 2002; Guo et al., 2015). Expectations for success and the value a student associates with this are assumed to directly influence performance, persistence, and choice in (academic-related) tasks.

Collectively, self-efficacy and outcome expectations shape the goals students set for themselves, based on both the expected outcome and the value they place on that outcome. These goals can be motivated by both the positive desire for a particular outcome and the desire to avoid negative outcomes (Eccles, 1994). Early in the development of these social psychology theories, researchers explored the influence of demographic—including age, gender, and race/ethnicity—and individual differences on the development of self-efficacy, outcome expectations, and career interest (Fouad and Smith, 1996). For example, closer to career entry, these goals may be aligned with actual career choices, while earlier in a student's academic life, they may be manifested in broader, more abstracted career interests. Thus, it is important to recognize the dynamic, reciprocal nature of self-efficacy, expectancy outcomes and academic-career goals that emerge from these psychological states and evolve over time.

1.3. Related work

Prior research on the relationship between students' non-cognitive aspects of learning in other STEM areas such as science and mathematics has shown a strong relationship between learning experiences, attitudinal variables such as self-efficacy and outcome expectancy, and future intentions with regards to career pathways (e.g., Beal and Crockett, 2010; DeWitt et al., 2014; Sadler et al., 2012). Emerging research in K-12 CS education has similarly revealed important relationships between gender (Guzdial et al., 2012), membership in underrepresented groups (McKlin et al., 2019), and future career paths (Orton et al., 2016).

Separate literature synthesis efforts by both Fraillon et al. (2019) and Román-González et al. (2019) reaffirmed the need to continue working on developing instruments designed to measure the cognitive dimensions of computer science and computational thinking (CT) understanding, and the associated non-cognitive factors such as self-efficacy and outcome expectancy that influence these cognitive outcomes. One of the earlier instruments related to these non-cognitive factors include the CS Attitude Survey (Wiebe, Williams, Yang, & Miller, 2003), though McKlin et al. (2019) and others have noted this instrument's basis in somewhat outdated psychological models and the lack of construct validation work to link it to more modern attitudinal and motivational models. More current works on the development and validation of the CS-attitudes instrument were conducted with students from non-English speaking population, which can create issues of measurement error when translated to English. (e.g., Korkmaz et al., 2017). Korkmaz et al.'s (2017) Computational Thinking Scales (CTS) is a 29-item instrument that has individuals self-report on their agreement with statements related to five different dimensions of computational thinking, such as algorithmic thinking and creativity. The instrument was validated with undergraduate students without clear theoretical guidance from non-cognitive constructs such as self-efficacy or outcome expectancy, though many items could be construed as measuring these factors. A similar effort by Kukul et al. (2017) was undertaken with the Computer Programming Self-efficacy Scale (CPSES). This 31-item instrument was based on self-efficacy theory and validated with 12–14 year old Turkish students. As with Korkmaz's instrument, they developed items around a computational thinking framework, with the items targeting self-efficacy related to fine-grained CT concepts (e.g., "I know how to use the programming variables"). These two recent instruments both take an approach to structure item sets around CT/CS constructs and attain reliability through the development of multiple items related to each facet of these CT practices or concepts. The broad nature of the CT/CS frameworks used in their development results in multiple factors and fairly long instruments. Long instruments are often the result of attempting to cover all facets of a construct but can create their own reliability and validity problems resulting from logistical time pressures or respondent fatigue (Maloney et al., 2011).

The most recent work on a CS-related self-efficacy scale is the Computer Programming Self-Efficacy Scale (CPSES) developed by Tsai et al. (2019). CPSES consists of fifteen Likert-scale items that measure students' self-efficacy in five programming domains: logical thinking, algorithm, control, debugging, and cooperation. Thus, this instrument takes the same general approach as Korkmaz et al. (2017) and Kukul et al. (2017) did in developing multiple items across different CT/CS practices. Even though the Tsai et al. (2019) claim that their instrument is capable of measuring students' perceptions of their computational thinking skills, the instrument lacks a clear exposition of the psychological theory that underpins its development. Moreover, the wording of some of the Tsai et al.'s CPSES items shown in publication (e.g., I can make use of divisions to enhance programming efficiency, and I can figure out program procedures without a sample.) raises questions as to whether validation was undertaken in the same language as that of the publication. As Messick

(1995) notes, theoretical framing and wording precision are essential in the instrument development process, because it is central to ensuring the content validity of an instrument. Thus, translations of instruments need to be revalidated. Finally, while they state the instrument is for use with middle school students and older, the validation sample was college-aged students.

1.4. Current work

Based on this literature review, there is a clear need for a validated, short self-report instrument that captures key attitudinal dimensions that might be affected by CT/CS interventions. Literature also points to middle grades as a logical starting point for development of such an instrument. A fine-grained understanding of these CT/CS practices and concepts, and instead target students' attitudes towards computer science and programming more generally. Guided by prior research work on non-cognitive STEM factors in this grade range, developing item sets would be done around the psychological constructs of self-efficacy and outcome expectancy, both of which have a strong theoretical and empirical basis and have been utilized extensively in the study of student engagement and persistence in STEM-oriented academics and career pathways. Rather than attempt to measure fine-grained distinctions in a students' attitudinal orientation towards specific CS concepts (e.g., loops, variables), this new instrument would capture broader orientation towards CT/CS based on these two constructs, thus resulting in a relatively compact instrument. Our argument would be that students in this age range are unlikely to have formed distinct responses regarding specific CS concepts and that even if they did, it would be of little utility in guiding intervention design.

The current work combines two robust psychometric approaches, Classical Test Theory (CTT) and Item-Response Theory (IRT) Rasch which have been well documented and utilized in the psychometric research literature base. These two approaches, in turn, will be guided by well-established construct validity frameworks (AERA, APA & NCME, 2014; Messick, 1995) to guide the process. There are a limited number of studies in CS education research that utilize these validation methods (e.g., Werner et al., 2012; Zendler, 2019), particularly in the case of attitudinal instruments. We believe that our current work, utilizing foundational psychological theory and robust validation methods will result in an improved instrument for measuring CS-related attitudes for middle grade students. The following sections will report on the development of such an instrument.

2. Method

2.1. Item development and validation procedure

The initial ten items of our MG-CS Attitudes instrument were adapted from an existing, validated attitudinal self-report instrument, the Student Attitudes toward STEM (S-STEM) Survey (Friday Institute, 2012). More specifically, the items were adapted from the Engineering and Technology attitudes subscale of the instrument. The items were all in Likert-scale type rated on five-point scales, from 1 (strongly disagree) to 5 (strongly agree). This subscale was also based on the two constructs of interest: self-efficacy and outcome expectancy. The wording of the items from the selected subscale was modified by changing the attitudinal focus from general engineering and technology to computer science. Table 1 presents the two constructs in the MG-CS Attitudes along with operational definition of and items associated with each construct. The validation of the S-STEM instrument was conducted using a sample of over 15,000 public school students in the state of North Carolina, USA and is reported in Unfried, Faber, Stanhope, & Wiebe (2015). Psychometric tests, including exploratory and confirmatory factor analysis, showed that the attitude constructs and career interest items were valid and reliable (Cronbach's Alpha ranged from 0.83 to 0.92), and tests for measurement invariance demonstrated that the survey measures the same information in the same ways across students of different ages, races/ethnicities, and genders.

This validation study follows the approach proposed by Messick (1995) along with the American Educational Research Association (AERA), American Psychological Association (APA), and National Council on Measurement in Education (NCME) (2014) to validate our MG-CS Attitudes scale. Based on Messick, an instrument should have gone through several phases of validation before it can be used and considered as a validated instrument. In general, Messick emphasizes two types of validity: content and construct validity. Regarding content validity, researchers may ensure this aspect of validity by doing a thorough literature review or using a well-accepted theory to underpin the development of the items. In the current study, we used a well-accepted theory, the expectancy-value theory (Eccles and Wigfield, 2002; Wigfield and Eccles, 2000), as our content validity. Moreover, we reviewed the previous studies on CS-attitudes scale development in order to compare our work with prior efforts.

Messick (1995) divided construct validity into five different aspects: substantive, structural, external, generalizability, and consequential. The substantive aspect is the construct validity related to the consistency of participants' responses to every item on the instrument. In this study, the substantive aspect was examined by using several reliability values, including Cronbach's alpha, Rasch person (plausible-value) and separation (item) reliability. Also, we used cutoffs suggested by DeVellis (2017) to interpret the reliability values, in which values more than 0.70 are considered satisfactory. The use of many types of reliability were intended to address the concerns around Cronbach's alpha, especially related to the nature of ordinal scale (Cronbach & Shavelson, 2004; Peterson & Kim, 2013). IRT is able to convert the ordinal scale to a ratio/interval scale and thus more reliably compute the target statistics (Bond & Fox,

Table 1
Initial MG-CS attitudes.

Construct	Operational Definition	Item Code	Item
Self-efficacy	Students' beliefs about their ability to successfully achieve goals related to computer science and programming.	Item_1	I would like to create new computer programs.
		Item_3	I am good at building computer programs.
		Item_4	I am good at fixing computer programs.
		Item_10	I believe I can be successful in a career in programming.
Outcome expectancy	Students' beliefs in the anticipated outcomes that result from engaging in computer science and programming related activities or behaviors.	Item_2	If I learn programming, then I can improve things that people use every day.
		Item_5	I am interested in what makes computers work.
		Item_6	Designing computer programs will be important for my future work.
		Item_7	I am curious about how computer programs work.
		Item_8	I would like to use creativity and innovation in my future work.
		Item_9	When I combine math and science, I can invent more useful computer programs.

2015; Boone et al., 2014; for more detail). We still report Cronbach's alpha in our manuscript because it is still highly utilized in the literature and many readers may not be familiar with the reported person and item reliabilities computed by our IRT analysis.

The second component of construct validity is the structural aspect, which examines the number of latent constructs or factors underlying the instrument as well as the quality of the items occupying each latent construct. In this study, a combination of exploratory factor analysis (EFA), confirmatory factor analysis (CFA), and multidimensional Rasch modeling was used to investigate the structural aspect of the construct validity. Following best practice, the total sample was equally divided into two data sets. The first data set, called the testing data set, contained 331 students and was used to perform EFA. The results from the EFA were then compared to the theoretical framing we used when developing the instrument. CFA and multidimensional Rasch modeling were then performed to the remaining 332 students, called as training data set, to confirm the number of factors underlying the instrument. For the CFA, we used the cutoffs of fit indices suggested by Hu and Bentler (1999) and Schreiber et al. (2006) with $X^2/df < 3$, CFI >0.95, TLI > 0.95 and RMSEA < 0.08 or 0.06. As an additional metric, we used the value of 0.50 as the cutoff for average variance extracted (AVE, Hair et al., 2019). To evaluate the outcomes from the multidimensional Rasch analysis, we used the approach suggested by Adams and Wu (2010) to investigate the best model by looking at the lower chi-square, final deviance and Akaike Information Criterion (AIC). Moreover, the quality of the item was assessed through mean square (MNSQ) values, using the cutoffs suggested by Linacre (1994), which were between 0.60 and 1.40.

Third, the external aspect of construct validity, or criterion validity, investigates the correlation of the instrument under development with instruments measuring similar constructs (Messick, 1995). This external aspect may also examine the correlation of the instrument with outcomes or behaviors theoretically or empirically related to the constructs being measured. Thus, the current study also asked students to answer the science attitude subscale of the S-STEM (Unfried et al., 2015) because of its complementary academic area and theoretical similarity. We also calculated the correlation between students' scores on the MG-CS Attitudes instrument with their confidence in using a computer and their self-reported previous programming experience. This was informed by several studies that found a positive correlation between students' CS attitudes and these other two measures (e.g., Román-González et al., 2018; Rozell and Gardner, 2000). Also, among the 332 students in the training data set, 243 of them took a CS Concepts Inventory (Rachmatullah et al., 2020) as a cognitive assessment, and thus we were able to calculate the correlation between these two constructs as well.

The fourth aspect is generalizability, which assumes the fairness of the instrument, meaning that there is no group of respondents being privileged or disadvantaged in responding to some of the items. We performed Differential Item Functioning (DIF) to investigate this aspect. DIF gender and previous programming experiences were done for this study, given that gender (e.g., Cai et al., 2017) and prior programming experiences (e.g., Yukselturk and Altiok, 2017) are demographic factors that have shown significant differences in attitudinal responses to CT/CS. Cutoff values suggested by Boone et al. (2014), which are > 0.64 showing bias items, were used to evaluate our items.

Lastly, the consequential aspect of construct validity examines the intended and unintended consequences of interpretation of the respondents' scores after performing a statistical analysis on it. Given the scope of our work, we focused on one intended consequence of use of this instrument, the potential differences in CT/CS attitudes based on gender and prior experience. In this study, we compared students' MG-CS Attitudes scores based on their gender and previous programming experience to see if we came to the same results as previous published studies.

2.2. Sample and data collection

A total of 663 middle-grade students that attended schools in the Southeastern region of the U.S. participated in this study. More than half of the students were in the sixth grade (58%) and the remaining students were in either seventh (12%) or eighth (30%) grade. Reported gender distribution of the students was near equal, with 48% female and 43% male, with the remaining 9% of the students not responding to this question. The students were ethnically diverse: 30% White, 18% Black/African American, 15% Hispanic/Latinx, 6% Multiracial, 4% Asian, 2% Native American/American Indian, and 25% other or not identified. We also asked students to identify the extent of their prior programming experience. The response was bifurcated, with 42% of students reporting limited and 49% reporting high prior programming experience, the remaining 9% not responding.

2.3. Data analysis

For the EFA, the number of factors was decided through evaluating the Eigenvalues and parallel analysis. EFA was done using the "nFactors" version 2.3.3 package in R-software (Raîche and Magis, 2010). CFA was performed in IBM SPSS Amos version 25 (Arbuckle, 2017) and Multidimensional Rasch Analysis and DIF were run in ConQuest version 4.14.2 (Adams et al., 2015). Composite reliability was computed alongside CFA results to provide the information regarding estimation reliability (Raykov, 1997). The Pearson bivariate correlation test was run to investigate the correlation between MG-CS Attitudes, CS Concepts Inventory scores, and science attitude subscale of the S-STEM. Even though these scores were derived from ordinal scales that are not appropriate to run Pearson correlation tests, as noted earlier, Rasch modeling allows for the conversion of these data types to ratio-interval data that are more appropriate for parametric tests (see Baker & Kim, 2017; Boone et al., 2014). The Spearman correlation test was done for the correlation between CS-attitudes and the two variables with ordinal data (confidence with using a computer and previous programming experience). Additionally, a two-way ANOVA test was used to seek the differences in gender and programming experience as well as the interaction between gender and programming experience. All of the correlational and ANOVA analyses were done in IBM SPSS version 25 (IBM Corp, 2017).

3. Results

3.1. Structural aspect

The structural aspect of construct validity was done first, given that it would provide the basis for further analysis or investigation of other aspects of construct validity. Although we conceptualized the development of the items based on expectancy-value theory (Eccles and Wigfield, 2002; Wigfield and Eccles, 2000), and expected to see two factors, self-efficacy and outcome expectancy, empirical evidence of the existence of these two factors is needed. To do so, we first ran an EFA analysis to see the number of factors and the item loadings in every factor. Fig. 1 shows a scree plot computed from EFA. Based on Eigenvalues and parallel analysis, only one factor was detected in our data set with 51.6% of the total variance explained. Moreover, Table 2 shows the factor loadings and the uniqueness values for every item. It can be seen from Table 2 that most of the items were well-loaded into one factor with a range of loadings from 0.552 to 0.796.

As EFA functions as our preliminary investigation for the number of factors residing in our instrument, we then further investigated the number of constructs through multidimensional Rasch model and CFA. As previously mentioned, multidimensional Rasch model provides an alternative, robust method for factor identification. The training data set was fitted with these two approaches. Table 3 provides the results computed through multidimensional Rasch modeling. It can be seen that the two-factor model of the MG-CS Attitudes had lower X^2, final deviance and AIC than one-factor and the difference was significant ($\Delta X^2 = 161.69$, $\Delta df = 1$, $p < .001$). In contrast to the EFA, Rasch indicated that the two-factor model of the MG-CS Attitudes was the best model for our

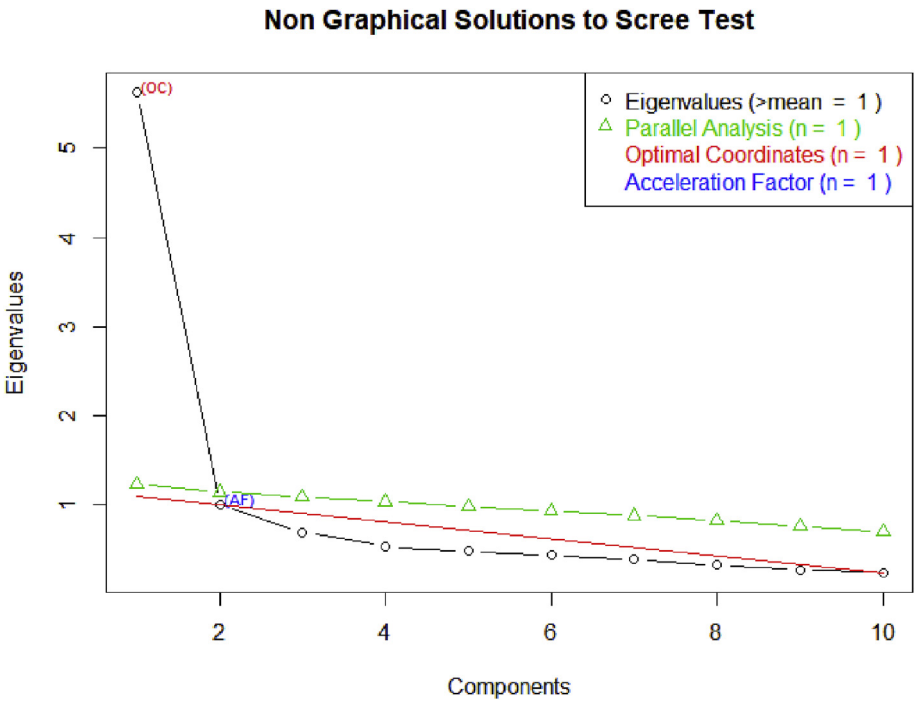

Fig. 1. The non-graphical solution to the scree test.

Table 2
Factor loadings and uniqueness values (*removed in later analysis due to misfit).

Item Code	Loading	Uniqueness	Item
Item_1	.796	.367	I would like to create new computer programs.
Item_2	.679	.539	If I learn programming, then I can improve things that people use every day.
Item_3	.678	.540	I am good at building computer programs.
Item_4*	.643	.587	I am good at fixing computer programs.
Item_5	.767	.411	I am interested in what makes computers work.
Item_6	.748	.441	Designing computer programs will be important for my future work.
Item_7	.797	.365	I am curious about how computer programs work.
Item_8	.552	.696	I would like to use creativity and innovation in my future work.
Item_9	.690	.524	When I combine math and science, I can invent more useful computer programs.
Item_10	.791	.374	I believe I can be successful in a career in programming.

instrument. A logical next step would be to investigate individual items for model fit.

The Rasch model also provided the indices to evaluate the quality of individual items by using MNSQ. In both the one-factor and two-factor models, Item 4 ("I am good at fixing computer programs.") was indicated as a misfitting item given that it had infit and outfit MNSQ of 1.83 and 1.88, respectively, in the one-factor model and 1.57 and 1.66 in the two-factor model. Thus, it was removed for further analysis. The Rasch two-factor model was rerun, and the new infit and outfit MNSQs for the remaining nine items are shown in Table 4. The revised model showed

Table 4
Rasch item fit indices and Cronbach's alpha if item deleted.

Dimension	Item Code	Measure	Infit MNSQ	Outfit MNSQ	Cronbach's Alpha if Item Deleted
Self-efficacy	Item_1	−0.27	0.91	0.90	.778
	Item_3	0.48	1.06	1.14	.801
	Item_10	−0.21	0.81	0.80	.753
Outcome expectancy	Item_2	−0.29	0.92	0.91	.832
	Item_5	0.30	1.03	1.01	.816
	Item_6	0.90	1.05	1.08	.835
	Item_7	0.18	0.95	0.95	.816
	Item_8	−0.88	1.33	1.38	.855
	Item_9	−0.21	0.98	0.97	.846

MNSQ values in the range of acceptable values (0.60–1.40), indicating all the items are well-fitted to the model. A Wright map showing students' abilities and item difficulties were also produced by the Rasch model, and it is visualized in Fig. 2.

We then compared the factorial structure of the instrument through CFA by comparing the one-factor model to the two-factor model, named Original and Revision 1 respectively. We found a better fit with the two-factor model based on the fit indices. However, the fit indices had not yet reached the cutoffs values. We then compared the Revision 1 model to the two-factor model without Item_4 (Revision 2), as the results from IRT-Rasch suggested. The comparison results are provided in Table 5. Based on Table 5, removing Item_4 from the model produced better fit indices. Analysis using the Amos software suggested a modification to connect the residual error of the Item_5 and Item_7. Reflecting on the wording of these two items, Item_5 and Item_7 had a similar emotional orientation, especially with the use of "interested" and "curious" as the wording, thus providing a parsimonious reason for connecting these

Table 3
Comparison between one-dimensional model and two-dimensional model of the MG-CS Attitudes.

Model	X^2	*df*	Final Deviance	AIC	Number of Parameters	Number of Misfitting Items
One-dimension	556.80***	9	8295.337	8323.337	14	1
Two-dimension	395.11***	8	8233.393	8265.393	16	1

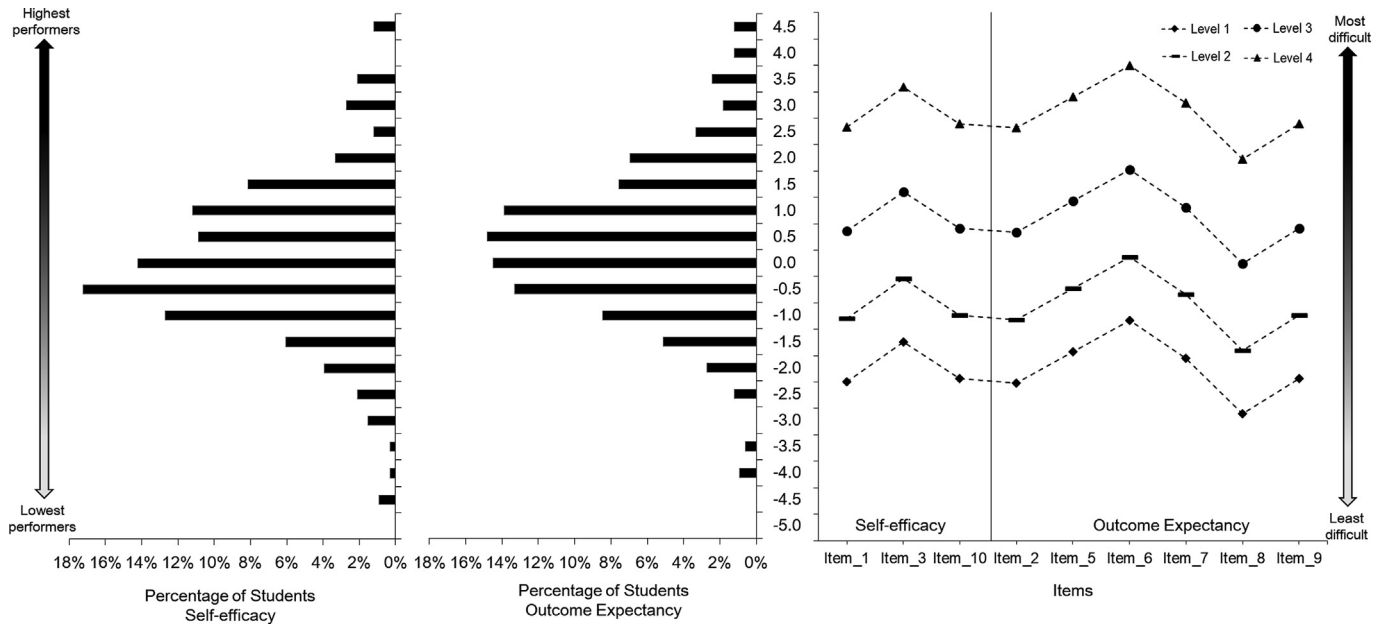

Fig. 2. Wright Map for the MG-CS Attitudes Scale.

Table 5
Comparison of CFA results among different models.

Indicator	Original	Revision 1	Revision 2	Revision 3
X^2	231.407	195.379	113.087	52.819
df	35	34	26	25
X^2/df (<3)	6.612	5.746	4.349	2.113
p-value	<.001	<.001	<.001	.001
CFI (>.95)	.892	.913	.946	.983
TLI (>.95)	.861	.860	.906	.969
RMSEA (<.08)	.132	.120	.101	.058
(90% CI)	(.116, .148)	(.104, .136)	(.082, .120)	(.036, .080)
$\Delta X^2(\Delta df)$	–	36.028(1)	82.292(8)	60.268(1)
p-value for ΔX^2	–	<.001	<.001	<.001

items. By making the two items correlated (Revision 3), the CFI, TLI, and RMSEA fit indices all improved to be within the range of acceptable values.

It can also be seen from Table 5, the Revision 3 model had better fit-indices than all types of one and two-factor models, most importantly with a confidence interval of the RMSEA that did not exceed the values of 0.08. The results of the CFA alternative model evaluations points to supporting the Rasch result recommending a two-factor solution as the basis for further analyses. Because Schreiber et al. (2006) argued that the chi-square test is possibly overly sensitive to large sample sizes, we ran chi-square test of differences (ΔX^2) between models to show how significant the improvement was from one model to another. Table 5 shows that each success improvement was significant, including the Revision 3 model. The average variance extracted (AVE) for the first factor (self--efficacy) was 0.63 and for the second factor (outcome expectancy) was 0.49. Even though the second factor had AVE less than the cutoff value (0.50), we believed that this is still acceptable given its proximity to the cutoff value. Moreover, the first factor's composite reliability value was 0.837, while the second factor was 0.852. Our final, recommended model, Revision 3, is visualized in Fig. 3.

3.2. Substantive aspect

The substantive aspect of construct validity assumes the stability of the students' responses to the items within each factor. Cronbach's alpha, Rasch person (plausible-value/PV) and separation reliabilities were used to investigate this aspect of construct validity. The Cronbach's alpha for self-efficacy was 0.840, and for outcome expectancy .858. Table 4 shows the "Cronbach's alpha if item deleted," indicating there was no noticeable improvement with the removal of any of the items. The Rasch person reliabilities for self-efficacy and outcome expectancy were .889 and .916 respectively. Finally, Rasch separation reliability for the two-factor model was 0.986. In summary, all of the reliability values exceeded the cutoff value of 0.70, indicating stable responses across samples and items using a two-factor model.

3.3. Construct aspect/criterion validity

Bivariate correlation coefficients were calculated to investigate the correlation between scores on MG-CS Attitudes scale and other measures indicated by theory or prior empirical work as expected to be related (e.g., Román-González et al., 2018; Rozell and Gardner, 2000). Expectancy-value theory predicts unique contributions but interrelatedness between self-efficacy and outcome expectancy (Wigfield and Eccles, 2000). In addition, studies of the S-STEM instrument (Unfried, Faber, Stanhope, & Wiebe, 2015), which the MG-CS Attitudes instrument was based on, show moderate correlation between the Science subscale and both the Mathematics and Engineering & Technology subscales. Thus we might expect moderate but unique correlations between the self-efficacy and outcome expectancy factors of the MG-CS Attitudes, and between MG-CS Attitudes scale and the S-STEM Science sub-scale. Expectancy-value theory would also predict a relationship of prior experience and self-efficacy of a common target area (Bandura, 1986; Wigfield and Eccles, 2000). Finally, one might assume that a broader confidence in using computers, and CS self-efficacy and outcome expectancy to be related.

The correlation between CS self-efficacy and outcome expectancy factors of the MG-CS Attitudes scale was $r = 0.815$. Though this correlation is high, it is still > 0.90, at which point the two factors might be considered as identical. Table 6 provides the results of further Pearson and Spearman correlation tests. Based on the results, CS self-efficacy and outcome expectancy were significantly correlated to science self-efficacy and outcome expectancy, even though the correlations were medium-weak (>0.25). Both MG-CS Attitudes factors were also significantly correlated with CS conceptual understanding score ($r = 0.392$ and $r = 0.395$) as well as confidence with using a computer ($r = 0.355$ and $r = 0.299$). The highest correlation was between CS self-efficacy and

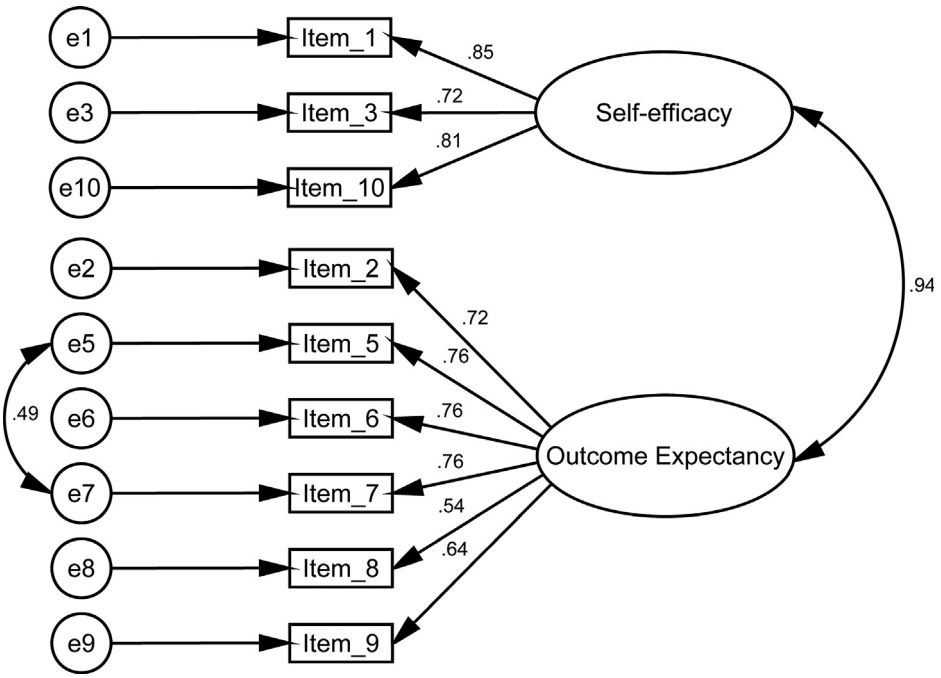

Fig. 3. Confirmatory factor analysis of the two-factor (Revision 3) final model.

Table 6
Correlation coefficients between the CS Self-efficacy and outcome expectancy factors of the MG-CS Attitudes scale and other related constructs (all correlations are significant at $\alpha = 0.01$).

Construct	CS Self-efficacy	CS Outcome Expectancy
Science Self-efficacy	.347	.304
Science Outcome Expectancy	.270	.311
CS Conceptual Understanding	.392	.395
Confidence Using Computer	.355	.299
Programming Experience	.527	.383

programming experience, which was $r = 0.527$, while between CS outcome expectancy and programming experience was more weakly correlated ($r = 0.383$) but still significant.

3.4. Generalizability aspect

DIF analysis was used to ensure that all the items behave in a consistent fashion to different populations likely to be targeted by this instrument. As noted in the prior section, the relationship between prior experience, and self-efficacy and outcome expectancy has been an active area of research (Usher and Pajares, 2008). Similarly, there is considerable interest in looking at the relationship of gender and CS-related attitudes (e.g., Knezek et al., 2015; Mindetbay et al., 2019). Table 7 shows the DIF contrasts for all the items in the two analyses of prior experience and gender. The results indicated that most of the DIF contrasts were less than the cutoff of 0.64, indicating that the items did not display bias based on gender or prior programming experience. It is important to note that while Item_8's DIF contrast for gender was technically over the cutoff (0.73), we consider this value acceptable given its proximity to the cutoff. However, use of the MG-CS Attitudes scale to study gender differences in CS outcome expectancy should take this into account.

3.5. Consequential aspect

Both prior experience and gender are considered to be consequential demographic factors with regards to attitudes towards CS (e.g., Alexandron et al., 2012; Lewis, 2010; Settle et al., 2015). First, a two-way

Table 7
DIF Contrasts for gender and previous experience.

Dimension	Item Code	DIF Gender	DIF Experience
Self-efficacy	Item_1	0.04	0.09
	Item_3	0.16	0.16
	Item_10	0.11	0.08
Outcome expectancy	Item_2	0.41	0.29
	Item_5	0.34	0.18
	Item_6	0.51	0.12
	Item_7	0.26	0.14
	Item_8	***0.73***	0.49
	Item_9	0.20	0.22

ANOVA test was used to investigate the interaction effect of gender and previous programming experience on both CS self-efficacy and outcome expectancy. With regards to self-efficacy, both in male and female groups, students with high experience ($M = 0.83$, $SD = 2.81$ and $M = -0.04$, $SD = 2.67$, respectively) had higher scores than those with limited previous programming experiences ($M = -0.13$, $SD = 2.36$ and $M = -1.60$, $SD = 2.80$, respectively). The interaction effect between gender and previous programming experience on CS self-efficacy was not significant ($F[1, 292] = 0.81$, $p = .370$, $\eta_p^2 = 0.003$). Similarly, for the CS outcome expectancy, both in both in male and female groups, students with high experience (M = 1.16, SD = 1.61 and M = 0.32, SD = 1.75, respectively) had higher scores than those with limited previous programming experiences (M = 0.67, SD = 1.88 and M = 0.06, SD = 1.72, respectively). Also, the interaction effect between gender and previous programming experience on CS outcome expectancy was not significant (F [1, 292] = 0.07, $p = .790$, $\eta_p^2 = 0.000$). The results are visualized in Fig. 4.

A two-way ANOVA test was then run again once more to the model by removing the interaction effect from the model to examine the main effects of gender and previous programming experiences on CS self-efficacy and outcome expectancy. For self-efficacy, it was found that males ($M = 0.54$, $SD = 2.70$) had higher scores than females ($M = -0.92$, $SD = 2.84$) and the difference was significant with medium effect size (F [1, 293] = 11.83, $p < .001$, $\eta_p^2 = 0.039$). Students with high programming experience ($M = 0.47$, $SD = 2.78$) had higher self-efficacy than those with

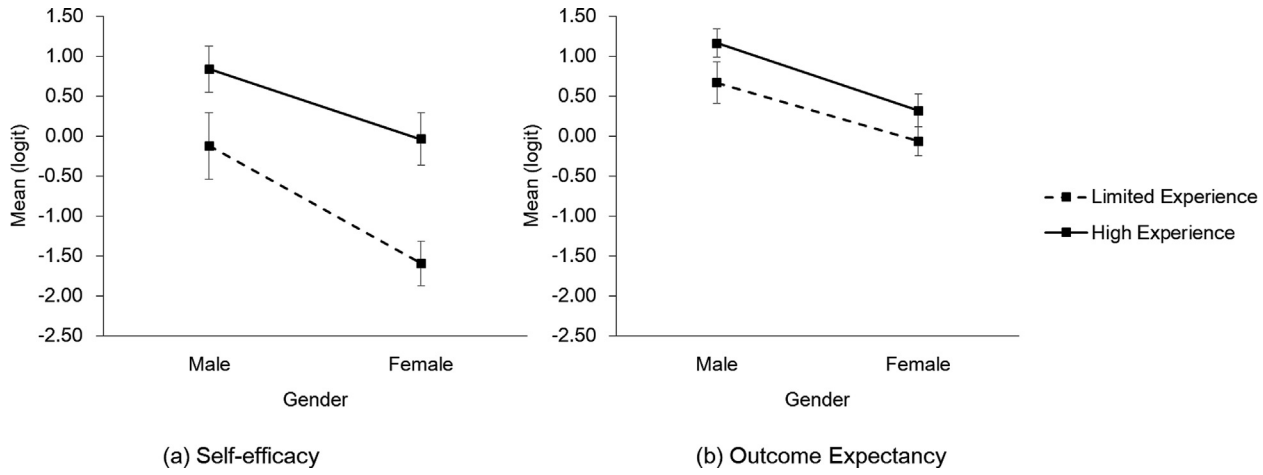

Fig. 4. Two-way effect of gender and programming experience on (a) CS self-efficacy and (b) outcome expectancy

limited programming experience ($M = -1.12$, $SD = 2.74$), and the difference was statistically significant with medium effect size (F [1, 293] < 15.57, $p = .001$, $\eta_p^2 = 0.050$). For outcome expectancy, in terms of gender, males ($M = 0.99$, $SD = 1.70$) were found to have a higher scores than females ($M = 0.10$, $SD = 1.74$), and this was statistically significant with a medium effect size (F [1, 293] = 14.80, $p < .001$, $\eta_p^2 = 0.048$). So to with the programming experience, students with high programming experience ($M = 0.81$, $SD = 1.72$) had significantly (F [1, 293] = 4.30, $p = .039$, $\eta_p^2 = 0.014$) higher scores than those with limited experience ($M = 0.17$, $SD = 1.80$).

4. Discussion

The validation of the MG-CS Attitudes scale was guided by well-established frameworks (AERA, APA & NCME, 2014; Messick, 1995) which provided sources of evidence across multiple aspects of validity. Our literature review and theoretical framing based on expectancy-value theory (Eccles and Wigfield, 2002; Wigfield and Eccles, 2000) provided our basis of content validity. A discussion of our findings related to five different aspects of construct validity are given below.

For structural validity, expectancy-value theory and prior related literature guided our exploration of both one and two-factor models. Whereas the EFA results indicated that a one-factor model might be the best fitting model, the multidimensional IRT Rasch and CFA ultimately showed a two-factor model provided the best fit for our data. This two-factor model consists of self-efficacy and outcome-expectancy, which is in alignment with what we had predicted in our theoretical framing (Eccles and Wigfield, 2002; Wigfield and Eccles, 2000). These results differ from the S-STEM Engineering and Technology subscale (Unfried et al., 2015) that was the basis of the MG-CS attitudes instrument, where only one factor was found for the subscale. One of the reasons for such discrepancy is that the validation process of the S-STEM was done in parallel with three disciplines—engineering and technology, science and mathematics—which might impact the factor loadings, resulting in factors based on disciplinary areas instead of the psychological dispositions underlying the scales. Based on Hofer (2000), this outcome may be because the factors based on disciplinary areas tend to have stronger influence on students, suggesting that student response is based more on disciplinary areas than on the more fine-grained psychological factors.

If we had only relied on EFA results, we would have likely settled on a one-factor model. Countering such an approach, Fabrigar et al. (1999) concluded that evaluating the number of factors based only on EFA results leads to poor interpretation of the latent psychological factors underlying surveys. Later studies by Piquero et al. (2000) and Lorenzo-Seva and Ferrando (2013) demonstrated how the use of several methods, specifically IRT, may better inform the researchers when deciding which statistical model they will apply to their instrument. Clearly, more studies on how students respond to and interpret the items in the MG-CS attitudes instrument is needed to clarify its dimensionality. The last structural aspect that is worth noting is the connection between the items containing "interest" and "curious" (Item 5 and 7). While these two words seem very similar, Litman and Silvia (2006) explain that curiosity is broader in interpretation than interest; where interest is part of curiosity but in a more specific way. Reflecting on the difficulty (to agree with) level provided in Table 3, the curiosity item was easier (0.18 logit) to agree with by students compared to the interest item (0.30 logit). Given these differences, we believe that both items are worth retaining.

Substantive validity for the two-factor model was provided by IRT-Rasch, demonstrating strong values for Cronbach's alpha (0.840 and 0.858 for self-efficacy and outcome expectancy, respectively), Rasch person (0.889 and 0.916) and separation (0.986) reliability, all exceeding the cutoff value of 0.70. In addition, IRT-Rasch showed that the Item_4 of the original instrument was a misfit item based on its MNSQ values. CFA results supported this item removal, with a significant improvement in the model with the misfit item removed. The misfitness might also come from the wording, especially regarding "fixing." The problematic use of "fixing" a computer or program was also raised by Grover et al. (2014) who found this term was not appropriate for middle school students with regards to computer programming.

Criterion validity was demonstrated with appropriately valued correlations with science self-efficacy, CS conceptual understanding and confidence with using a computer. Theory and prior empirical work predicted the demonstrated significant but moderate levels of correlation with all of these alternate measures. Perhaps, as expected, the strongest of these correlations ($r = 0.527$) was seen between Programming Experience and CS Self-efficacy. Generalizability was demonstrated using DIF analysis with gender and previous programming experience. Public policy and associated research (e.g., Stoilescu and Egodawatte, 2010; Tsan, Boyer, & Lynch, 2016) has shown a strong interest in studying the efficacy of interventions addressing gender inequality. For that reason, it was important to demonstrate that psychometrically, our instrument functioned similarly with boys and girls. Similarly, prior research has demonstrated (e.g., Alexandron et al., 2012; Lewis, 2010) a wide range of prior programming experience among middle grades students, so it was important for us to demonstrate similar uniform item functioning for this individual difference. These same two variables were used to demonstrate consequential validity. Our analysis showed a significant relationship between prior programming experience and self-efficacy, supporting a central tenet of the relationship of experience with a subject area and self-efficacy (Bandura, 1986). Similarly, the overall higher CS self-efficacy for boys than girls parallels similar findings by other researchers (e.g., Beyer, 2014; Huang, 2013).

5. Limitations

Because validation of an instrument is inherently an ongoing, iterative process, we wanted to reflect what we feel are some of the limitations of this initial analysis and possible areas for future work. First, while we had a sample that was more than adequate for the statistical analyses conducted, and the sample had a relatively representative racial and ethnic diversity based on the U.S. population, it still falls quite short of being a true sample of either the U.S. or international population of middle grades (ages 11–13) students. This sample was taken from a region in a single state in the U.S. and therefore does not represent the true cultural diversity of the U.S. or internationally. In addition, this instrument currently only exists in the English language. We welcome other researchers using and studying this instrument in other locales and adding their findings to the research base. Second, the process of item testing and refinement has left the instrument with only three self-efficacy items. While this clearly helps meet the goal of a compact instrument, and with only three items that instrument continued to perform well psychometrically, it does fall short of the general heuristic of five items per construct for this type of instrument (Yong and Pearce, 2013). Finally, the high correlation coefficient between the two scales motivates us and other researchers to continue exploring the relationship between expectancy-value theory and its measurement. Based on our findings, we recommend treating it as a two-dimension construct, but further research should be conducted to explore the utility of treating it as a one-dimensional construct.

6. Conclusion

In conclusion, we feel that the MG-CS Attitudes instrument provides researchers with a compact, reliable, and well-validated instrument for measuring self-efficacy and outcome expectancy—two well accepted psychological constructs used in educational research. In addition, it addresses some of the short-comings of other attitudinal instruments by being well grounded in psychological theory and designing items with computer science as a singular target area rather than having middle grades students attempt to discern self-efficacy on fine-grained target concepts (e.g., debugging; Tsai et al., 2019) or on targets that pertain to more general practices (e.g., problem-solving; Korkmaz et al., 2017).

Further work should explore designing and testing new self-efficacy items that could be added to the instrument. Finally, this work has only added to the ongoing debate as to whether self-efficacy and outcome-expectancy represent one or two unique constructs in a self-report instrument such as this. Such questions have been debated not only with instruments designed for younger students, but also those designed for teachers targeting their attitudes towards teaching STEM subjects (e.g., Lekhu, 2013). Further work needs to be conducted to explore this interesting and important element of psychological theory as it relates to CS and other academic areas.

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.

Acknowledgments

This material is based upon work supported by the United States National Science Foundation under Grant Nos. DRL1640141 and CNS-1138497. 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.

References

Adams, R. J., Wu, M. L., & Wilson, M. R. (2015). *ACER ConQuest: Generalised Item Response Modelling Software*. Camberwell, Victoria: Australian Council for Educational Research [Computer software]. Version 4.

Adams, R., & Wu, M. (2010). *Notes and Tutorial ConQuest: Multidimensional Model*. Retrieved from https://www.acer.org/conquest/notes-tutorials.

Alexandron, G., Armoni, M., Gordon, M., & Harel, D. (2012). The effect of previous programming experience on the learning of scenario-based programming. In *Proceedings of the 12th Koli Calling International Conference on Computing Education Research - Koli Calling '12* (pp. 151–159). New York, New York, USA: ACM Press. https://doi.org/10.1145/2401796.2401821.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for Educational and Psychological Testing*. Washington, DC: AERA.

Arbuckle, J. L. (2017). *Amos (Version 25.0) [Computer Program]*. Chicago: SPSS.

Baker, F., B., & Kim, S. (2017). *The basic of item response theory using R*. Springer, Cham.

Bandura, A. (1986). *Social Foundations of Thought and Action: A Social Cognitive Theory*. Englewood Cliffs, NJ: Prentice-Hall.

Basawapatna, A. R., Koh, K. H., & Repenning, A. (2010). Using scalable game design to teach computer science from middle school to graduate school. In *Proceedings of the 15th Annual Conference on Innovation and Technology in Computer Science Education* (pp. 224–228). ACM.

Beal, S. J., & Crockett, L. J. (2010). Adolescents' occupational and educational aspirations and expectations: links to high school activities and adult educational attainment. *Dev. Psychol., 46*(1), 258–265. https://doi.org/10.1037/a0017416.

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. *Comput. Sci. Educ., 24*(2–3), 153–192. https://doi.org/10.1080/08993408.2014.963363.

Bond, T., G., & Fox, C., M. (2015). *Applying the Rasch model: fundamental measurement in the human sciences, 3rd ed*. New York, NY: Routledge.

Boone, W. J., Staver, J. R., & Yale, M. S. (2014). *Rasch Analysis in the Human Sciences*. Dordrecht, Netherlands: Springer.

Cai, Z., Fan, X., & Du, J. (2017). Gender and attitudes toward technology use: a meta-analysis. *Comput. Educ., 105*, 1–13. https://doi.org/10.1016/j.compedu.2016.11.003.

Chen, G., Gully, S. M., & Eden, D. (2001). Validation of a new general self-efficacy scale. *Organ. Res. Methods, 4*(1), 62–83. https://doi.org/10.1177/109442810141004.

Cronbach, L., J., & Shavelson, R., J. (2004). My current thoughts on coefficient alpha and successor procedures. *Educational and Psychological Measurement, 64*(3), 391–418. https://doi.org/10.1177/0013164404266386.

Deci, E. L., & Ryan, R. M. (1985). *Intrinsic Motivation and Self-Determination in Human Behavior*. New York, NY: Plenum.

Denner, J., Werner, L., & Ortiz, E. (2012). Computer games created by middle school girls: can they be used to measure understanding of computer science concepts? *Comput. Educ., 58*(1), 240–249. https://doi.org/10.1016/j.compedu.2011.08.006.

DeVellis, R. F. (2017). *Scale Development: Theory and Applications* (fourth ed.). Los Angeles, CA: Sage.

DeWitt, J., Archer, L., & Osborne, J. (2014). Science-related aspirations across the primary–secondary divide: evidence from two surveys in England. *Int. J. Sci. Educ., 36*(10), 1609–1629. https://doi.org/10.1080/09500693.2013.871659.

Eccles, J. S. (1994). Understanding women's educational and occupational choices: applying the Eccles et al. Model of achievement-related choices. *Psychol. Women Q., 18*(4), 585–609. https://doi.org/10.1111/j.1471-6402.1994.tb01049.x.

Eccles, J. S., & Wigfield, A. (2002). Motivational beliefs, values, and goals. *Annu. Rev. Psychol., 53*, 109–132. https://doi.org/10.1146/annurev.psych.53.100901.135153.

Edwards, M. (2011). Algorithmic composition: computational thinking in music. *Commun. ACM, 54*(7), 58–67.

Else-Quest, N. M., Mineo, C. C., & Higgins, A. (2013). Math and science attitudes and achievement at the intersection of gender and ethnicity. *Psychol. Women Q., 37*(3), 293–309. https://doi.org/10.1177/0361684313480694.

Fabrigar, L. R., Wegener, D. T., MacCallum, R. C., & Strahan, E. J. (1999). Evaluating the use of exploratory factor analysis in psychological research. *Psychol. Methods, 4*(3), 272–299.

Fouad, N. A., & Smith, P. L. (1996). A test of a social cognitive model for middle school students: math and science. *J. Counsel. Psychol., 43*(3), 338–346. https://doi.org/10.1037/0022-0167.43.3.338.

Fraillon, J., Ainley, J., Schulz, W., Duckworth, D., & Friedman, T. (2019). *International Computer and Information Literacy Study 2018: Assessment Framework*. Amsterdam, Netherlands: International Association for the Evaluation of Educational Achievement (IEA).

Friday Institute for Educational Innovation. (2012). *Middle and High School STEM-Student Survey*. Raleigh, NC: Author.

Google, & Gallup. (2016). *Diversity Gaps in Computer Science: Exploring the Underrepresentation of Girls, Blacks and Hispanics*. Retrieved from http://goo.gl/PG34aH.

Grover, S., Pea, R., & Cooper, S. (2014). March). Remedying misperceptions of computer science among middle school students. In *Proceedings of the 45th ACM Technical Symposium on Computer Science Education* (pp. 343–348). ACM.

Grover, S., Pea, R., & Cooper, S. (2016). February). Factors influencing computer science learning in middle school. In *Paper Presented at the Proceedings of the 47th ACM Technical Symposium on Computer Science Education* (pp. 552–557). ACM.

Groves, R. M., Fowler, F. J., Couper, M. P., Lepkowski, J. M., Singer, E., & Tourangeau, R. (2009). *Survey Methodology* (second ed.). Hoboken, NJ: John Wiley & Sons, Inc.

Guo, J., Marsh, H. W., Morin, A. J., Parker, P. D., & Kaur, G. (2015). Directionality of the associations of high school expectancy-value, aspirations, and attainment: a longitudinal study. *Am. Educ. Res. J., 52*(2), 371–402. https://doi.org/10.3102/0002831214565786.

Guzdial, M., Ericson, B. J., McKlin, T., & Engelman, S. (2012). A statewide survey on computing education pathways and influences: factors in broadening participation in computing. In *Proceedings of the 9th Annual International Conference on International Computing Education Research* (pp. 143–150). ACM.

Hair, J. F., Black, W. C., Babin, B. J., & Anderson, R. E. (2019). *Multivariate Data Analysis* (Eight edition). Hampshire, UK: Cengage Learning.

Hofer, B. K. (2000). Dimensionality and disciplinary differences in personal epistemology. *Contemp. Educ. Psychol., 25*(4), 378–405. https://doi.org/10.1006/ceps.1999.1026.

Hu, L. T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: conventional criteria versus new alternatives. *Struct. Equ. Model.: A Multidiscip. J., 6*(1), 1–55. https://doi.org/10.1080/10705519909540118.

Huang, C. (2013). Gender differences in academic self-efficacy: a meta-analysis. *Eur. J. Psychol. Educ., 28*(1), 1–35. https://doi.org/10.1007/s10212-011-0097-y.

IBM Corp. (2017). *IBM SPSS Statistics For Windows, Version 25.0 [Computer Software]*. Armonk, NY: IBM Corp.

Knezek, G., Christensen, R., Tyler-Wood, T., & Gibson, D. (2015). Gender differences in conceptualizations of STEM career interest: complementary perspectives from data mining, multivariate data analysis and multidimensional scaling. *J. STEM Educ., 16*(4), 13–19. https://www.learntechlib.org/p/171343/.

Korkmaz, Ö., Çakir, R., & Özden, M. Y. (2017). A validity and reliability study of the computational thinking scales (CTS). *Comput. Hum. Behav., 72*, 558–569. https://doi.org/10.1016/j.chb.2017.01.005.

Kukul, V., Gökçearslan, Ş., & Günbatar, M. S. (2017). Computer programming self-efficacy scale (CPSES) for secondary school students: development, validation and reliability. *Eğitim Teknol.Kuram Uygulama [Educ. Technol. Theory Practice], 7*(1), 158–179. Retrieved from http://dergipark.ulakbim.gov.tr/etku/article/view/5000195912.

Lekhu, M. A. (2013). Relationship between self-efficacy beliefs of science teachers and their confidence in content knowledge. *J. Psychol. Afr., 23*(1), 109–112. https://doi.org/10.1080/14330237.2013.10820602.

Levitan, M. (2018, December 17). *Report: Women and Minorities Continue to Be Underrepresented in Computer Science. Diverse Issues in Higher Education*. Retrieved from https://diverseeducation.com/article/134588/.

Lewis, C. M. (2010). How programming environment shapes perception, learning and goals. In *Proceedings of the 41st ACM Technical Symposium on Computer Science Education - SIGCSE'10* (pp. 346–350). New York, NY: ACM Press. https://doi.org/10.1145/1734263.1734383.

Linacre, J. M. (1994). Sample size and item calibration stability. *Rasch Meas. Trans., 7*(4), 328.

Litman, J. A., & Silvia, P. J. (2006). The latent structure of trait curiosity: evidence for interest and deprivation curiosity dimensions. *J. Pers. Assess., 86*(3), 318–328. https://doi.org/10.1207/s15327752jpa8603_07.

Lorenzo-Seva, U., & Ferrando, P. J. (2013). Factor 9.2: a comprehensive program for fitting exploratory and semiconfirmatory factor analysis and IRT models. *Appl. Psychol. Meas., 37*(6), 497–498. https://doi.org/10.1177/0146621613487794.

Maloney, P., Grawitch, M. J., & Barber, L. K. (2011). Strategic item selection to reduce survey length: reduction in validity? *Consult. Psychol. J. Pract. Res., 63*(3), 162–175. https://doi.org/10.1037/a0025604.

McKlin, T., Wanzer, D., Lee, T., Magerko, B., Edwards, D., Grossman, S., & Freeman, J. (2019, February). Implementing EarSketch: connecting classroom implementation to student outcomes. In *Proceedings of the 50th ACM Technical Symposium on Computer Science Education* (pp. 634–640). ACM.

Messick, S. (1995). Validity of psychological assessment: validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *Am. Psychol., 50*(9), 741–749.

Mindetbay, Y., Bokhove, C., & Woollard, J. (2019). What is the relationship between students' computational thinking performance and school achievement? *Int. J. Comput. Sci. Eng. Syst.*, 1–13. https://doi.org/10.21585/ijcses.v0i0.45.

Orton, K., Weintrop, D., Beheshti, E., Horn, M., Jona, K., & Wilensky, U. (2016). Bringing computational thinking into high school mathematics and science classrooms. In *Paper Presented at the International Conference on the Learning Sciences ICLS, Singapore*.

Osborne, J., Simon, S., & Collins, S. (2003). Attitudes towards science: a review of the literature and its implications. *Int. J. Sci. Educ., 25*(9), 1049–1079. https://doi.org/10.1080/0950069032000032199.

Peterson, R., A., & Kim, Y. (2013). On the relationship between coefficient alpha and composite reliability. *Journal of Applied Psychology, 98*(1), 134–138. https://doi.org/10.1037/a0030767.

Piquero, A. R., MacIntosh, R., & Hickman, M. (2000). Does self-control affect survey response? Applying exploratory, confirmatory, and item response theory analysis to Grasmick et al.'s self-control scale. *Criminology, 38*(3), 897–930. https://doi.org/10.1111/j.1745-9125.2000.tb00910.x.

Porter, A. C. (2006). Curriculum assessment. In J. C. Green, G. Camill, & P. B. Elmore (Eds.), *Complementary Methods for Research in Education* (pp. 141–160). Washington, DC: American Educational Research Association.

Rachmatullah, A., Akram, B., Boulden, D., Mott, B., Boyer, K., Lester, J., & Wiebe, E. (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), 1–11.

Raîche, G., & Magis, D. (2010). *nFactors: Parallel Analysis and Non Graphical Solutions to the Cattell's Scree Test* [R package Version 2.3.1.]. Retrieved from cran.r-project.org/web/packages/nFactors/index.html.

Raykov, T. (1997). Estimation of composite reliability for congeneric measures. *Appl. Psychol. Meas., 21*(2), 173–184. https://doi.org/10.1177/01466216970212006.

Román-González, M., Moreno-León, J., & Robles, G. (2019). Combining assessment tools for a comprehensive evaluation of computational thinking interventions. In S. C. Kong, & H. Abelson (Eds.), *Computational Thinking Education* (pp. 79–98). Gateway East, Singapore: Springer Nature Springer.

Román-González, M., Pérez-González, J. C., Moreno-León, J., & Robles, G. (2018). Extending the nomological network of computational thinking with non-cognitive factors. *Comput. Hum. Behav., 80*, 441–459. https://doi.org/10.1016/j.chb.2017.09.030.

Rozell, E. J., & Gardner, W. L., III (2000). Cognitive, motivation, and affective processes associated with computer-related performance: a path analysis. *Comput. Hum. Behav., 16*(2), 199–222. https://doi.org/10.1016/S0747-5632(99)00054-0.

Sadler, P. M., Sonnert, G., Hazari, Z., & Tai, R. (2012). Stability and volatility of STEM career interest in high school: a gender study. *Sci. Educ., 96*(3), 411–427. https://doi.org/10.1002/sce.21007.

Schreiber, J. B., Nora, A., Stage, F. K., Barlow, E. A., & King, J. (2006). Reporting structural equation modeling and confirmatory factor analysis results: a review. *J. Educ. Res., 99*(6), 323–338. https://doi.org/10.3200/JOER.99.6.323-338.

Settle, A., Lalor, J., & Steinbach, T. (2015). Reconsidering the impact of CS1 on novice attitudes. In *Paper Presented at the Proceedings of the 46th ACM Technical Symposium on Computer Science Education*.

Sharek, D., & Wiebe, E. (2014). Measuring video game engagement through the cognitive and affective dimensions. *Simulation & Gaming, 45*(4–5), 569–592. https://doi.org/10.1177/1046878114554176.

Shin, S., Rachmatullah, A., Roshayanti, F., Ha, M., & Lee, J. K. (2018). Career motivation of secondary students in STEM: a cross-cultural study between Korea and Indonesia. *Int. J. Educ. Vocat. Guid., 18*(2), 203–231. https://doi.org/10.1007/s10775-017-9355-0.

Stanton, J., Goldsmith, L., Adrion, W. R., Dunton, S., Hendrickson, K. A., Peterfreund, A., & Zinth, J. D. (2017). *State of the States Landscape Report: State-Level Policies Supporting Equitable K–12 Computer Science Education*. Retrieved from https://www.edc.org/state-states-landscape-report-state-level-policies-supporting-equitable-k-12-computer-science.

Stoilescu, D., & Egodawatte, G. (2010). Gender differences in the use of computers, programming, and peer interactions in computer science classrooms. *Comput. Sci. Educ., 20*(4), 283–300. https://doi.org/10.1080/08993408.2010.527691.

Summers, R., & Abd-El-Khalick, F. (2018). Development and validation of an instrument to assess student attitudes toward science across grades 5 through 10. *J. Res. Sci. Teach., 55*(2), 172–205. https://doi.org/10.1002/tea.21416.

Tourangeau, R., Rips, L. J., & Rasinski, K. (2000). *The Psychology of Survey Response*. Cambridge, UK: Cambridge University Press.

Tsai, M. J., Wang, C. Y., & Hsu, P. F. (2019). Developing the computer programming self-efficacy scale for computer literacy education. *J. Educ. Comput. Res., 56*(8), 1345–1360. https://doi.org/10.1177/0735633117746747.

Tsan, J., Boyer, K. E., & Lynch, C. F. (2016, February). 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.

Usher, E. L., & Pajares, F. (2008). Sources of self-efficacy in school: critical review of the literature and future directions. *Rev. Educ. Res., 78*(4), 751–796.

Werner, L., Denner, J., Campe, S., & Kawamoto, D. C. (2012). February). The fairy performance assessment: measuring computational thinking in middle school. In *Proceedings of the 43rd ACM Technical Symposium on Computer Science Education* (pp. 215–220). ACM.

Wiebe, E., Williams, L. A., Yang, K., & Miller, C. S. (2003). *Computer science attitude survey*. North Carolina State University. Dept. of Computer Science.

Wigfield, A., & Eccles, J. S. (2000). Expectancy–value theory of achievement motivation. *Contemp. Educ. Psychol., 25*(1), 68–81. https://doi.org/10.1006/ceps.1999.1015.

Wilkerson-Jerde, M. H., Gravel, B. E., & Macrander, C. A. (2015). Exploring shifts in middle school learners' modeling activity while generating drawings, animations, and computational simulations of molecular diffusion. *J. Sci. Educ. Technol., 24*(2–3), 396–415. https://doi.org/10.1007/s10956-014-9497-5.

Yong, A. G., & Pearce, S. (2013). A beginner's guide to factor analysis: focusing on exploratory factor analysis. *Tutorials in Quant. Methods Psychol., 9*(2), 79–94. https://doi.org/10.20982/tqmp.09.2.p079.

Yukselturk, E., & Altiok, S. (2017). An investigation of the effects of programming with Scratch on the preservice IT teachers' self-efficacy perceptions and attitudes towards computer programming. *Br. J. Educ. Technol., 48*(3), 789–801. https://doi.org/10.1111/bjet.12453.

Zendler, A. (2019). cpm. 4. CSE/IRT: compact process model for measuring competences in computer science education based on IRT models. *Educ. Inf. Technol., 24*(1), 843–884.