Technical Report 1

The Development of Computer Science Concepts Inventory for Middle School Students: Preliminary Results

January 2019

Arif Rachmatullah

Danielle Boulden

Eric Wiebe

The William and Ida Friday Institute for Educational Innovation
North Carolina State University
Raleigh, N.C.

NC STATE
College of Education

Background, Purposes and the Intended Uses of the Assessment

The CS Concepts Inventory is intended to measure students' understanding of the four core concepts of CS—variables, conditionals, loops and algorithms— taught at the middle school level. Additionally, we incorporated the concepts of debugging, comprehension and development into the assessment. The assessment was guided by a conceptual framework informed by a Focal Knowledge, Skills and Abilities—FKSAs framework developed by Grover and Basu (2017), the K-12 CS Framework (K–12 Computer Science Framework, 2016) and the Computer Science Teachers Association (CSTA) Standards (CSTA, 2017). The assessment utilizes elements from a block-based programming environment as the context for every question, based on findings that suggest learners, especially novice ones, experience less conceptual and cognitive difficulties using these tools (e.g., Grover, Pea & Cooper, 2015; Robins, Rountree, & Rountree, 2003).

The current version of the CS Concepts Inventory was written for students in grades sixth through eighth. We believe the assessment can be used in either longitudinal contexts, such as in pre-intervention-post design or as a sole administration such as a pre or post assessment only. The assessment is designed in multiple-choice format with four to five answer choices. There are 22 items in total on the assessment. We suggest giving students approximately 40-45 minutes to complete the administration.

The Process of the Development of the Assessment

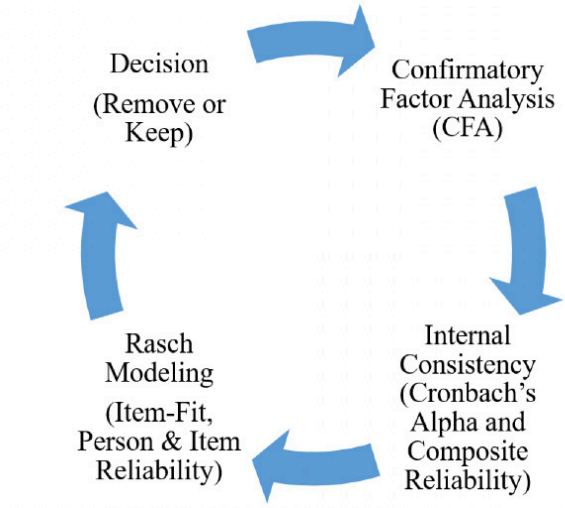
Some of the items utilized for the assessment were adapted from Weintrop and Wilensky (2015) and work done in the labs of Thomas Price and Tiffany Barnes (Department of Computer Science, NC State University). The remaining items were developed based on the FKSAs, K-12 CS Framework, and CSTA Standards (noted above). There were two pilot studies conducted on the assessment that tested two different versions of the instrument. The original draft version that was used in the first pilot study consisted of 20 items and was administered to 22 middle school students. The results of this first study was used to determine the spread of item difficulty level, as well as readability aspects of the assessment. The results of the first pilot study showed that the items were concentrated in the medium and hard level of difficulty. Thus, in the second pilot study, we added seven additional items with the hope of having a broader range of item difficulty levels, especially in the lower level of difficulty. For each pilot study, we asked three undergraduate and graduate students that were considered as novices in computer science, to give feedback regarding the clarity and readability of the assessment items and directions.

The revised version of the assessment consisted of 27 items and used in a second pilot study that was administered to 245 middle school students. Confirmatory Factor Analysis (CFA), Cronbach's alpha, Composite Reliability – CR (Raykov, 1997) and Rasch modeling analysis were used to assess the reliability and validity this revised version of the assessment. Those analyses yielded results indicating a satisfactory level of model characteristics for assessment in research settings. This second round of analysis resulted in a model that consisted of 22 final items.

Evidence of Construct Validity and Reliability

To collect the evidence of validity and reliability, we combined the Classical Test Theory (CTT) method and the Item Response Theory (IRT) method. We first ran a CFA to confirm that the assessment is unidimensional and measures one latent trait – students' understanding of CS concepts. Next, we calculated Cronbach's alpha and performed CR tests to examine the internal consistency of the assessment. Finally, we ran Rasch modeling analysis to test the validity and reliability of the assessment, given that Rasch modeling analysis provides more robust results of validation. Based on these results, decisions were made on which items to retain and which to remove. Items were removed from the model when they were not reaching statistical significance ($p >$ .05) on the resulting CFA. By removing the items, it increased the value of Cronbach's alpha. Items were also removed when the values of outfit and infit MNSQ were beyond the acceptable range of 0.70 – 1.30 (Wright & Linacre, 1994). All the analyses were done using WINSTEP 4.0.1 and Stata 15. Figure 1 is a diagram depicting the procedure used to validate the assessment.

Figure 1. *The Procedure of Validation*



After four misfitting items were removed from the model, we obtained the result of CFA with $X^2/df = 1.33$, $p =$ .001, RMSEA = .039 (90%CI = .026, .051), CFI = .863, TLI = .849 and SRMR = .059. Even though the CFI and TLI values did not meet the acceptable value of > 0.95 (Hu & Bentler, 1999), we had X^2/df and RMSEA values that met the cutoff, which are < 2 (Tabachnick & Fidell, 2007) and < .06 (Hu & Bentler, 1999), respectively. Given that we are still in the process refining this assessment, and CFA is dependent upon sample size, we believe this result is expected to change with future administrations of the instrument as the sample size increases. Moreover, Diamantopoulos and Siguaw (2000) argue that RMSEA is an important aspect of CFA because it detects the lack of fit between the obtained data and the model.

Even though we consider all the results from the CTT method, in this case, CFA results, we rely more heavily on the results computed through Rasch analysis. One of the reasons is Rasch analysis is deemed to be sample-independent and thus does not depend on sample size (Bond & Fox, 2013; Boone, Staver & Yale, 2013).

Moreover, Rasch analysis has an assumption that higher ability students have a higher probability to correctly answer both more difficult items and easier items than lower ability students (Bond & Fox, 2013; Boone, Staver & Yale, 2013), which is the primary role of an assessment. This assumption is reflected in MNSQ values, when the values are outside the standards mentioned above, the item does not behave as it should be. Table 1 shows all the MNSQ values, and they are in the range of acceptable values except outfit MNSQ value for Item_4_Loo3 which is 0.69. We believe this value is still acceptable due to its close proximity to the cutoff.

Regarding the internal consistency of the assessment, all of the reliability values computed through both CTT (α and CR) and Rasch methods (person and item reliabilities) were acceptable which is > .70 (DeVellis, 2003). The final analyses yielded the following values: Cronbach's alpha = 0.784, CR = 0.828, and Rasch person and item reliabilities were 0.76 and 0.94, respectively.

Table 1. *The Results of Cronbach's α and Rasch Modeling Analysis*

Item	α If Item Deleted	Measure	Infit MNSQ	Outfit MNSQ	PTMA
Item_1_Var1	0.777	0.17	1.02	1.00	0.45
Item_3_Loo1	0.777	0.93	1.00	1.05	0.46
Item_4_Loo2	0.767	-1.09	0.82	0.69	0.39
Item_5_Alg1	0.785	-0.05	1.18	1.25	0.44
Item_6_Var2	0.787	-0.07	1.19	1.29	0.44
Item_9_Con4	0.772	-0.46	0.97	0.92	0.43
Item_10_Loo3	0.773	-0.84	0.91	0.86	0.40
Item_12_Var3	0.782	-0.52	1.08	1.10	0.42
Item_13_Alg2	0.777	0.31	1.01	1.04	0.45
Item_14_Var4	0.769	-0.03	0.95	0.96	0.44
Item_16_Alg3	0.779	0.16	1.05	1.05	0.45
Item_17_Var5	0.765	-0.07	0.84	0.81	0.44
Item_18_Var6	0.772	0.32	0.95	0.94	0.45
Item_19_Alg4	0.784	1.25	1.17	1.22	0.45
Item_20_Alg5	0.783	0.13	1.12	1.14	0.45
Item_21_Var7	0.774	0.22	0.98	0.97	0.44
Item_22_Var8	0.773	0.77	0.97	0.96	0.44
Item_23_Con5	0.781	-0.88	1.05	1.00	0.39
Item_24_Loo6	0.772	0.58	0.94	0.93	0.44
Item_25_Loo7	0.770	-0.16	0.91	0.90	0.43
Item_26_Alg6	0.773	0.76	0.95	0.94	0.44
Item_27_Alg7	0.776	-1.43	0.92	0.83	0.35

References

Bond, T. G., & Fox, C. M. (2013). *Applying the Rasch model: Fundamental measurement in the human sciences*. New York, NY: Psychology Press.

Computer Science Teachers Association (CSTA). (2017). *K-12 Computer Science Standards, Revised 2017*. Retrieved from https://www.csteachers.org/page/standards.

DeVellis, R. F. (2003). *Scale development: Theory and applications (2nd ed.)*. Thousand Oaks, CA: Sage.

Diamantopoulos, A. & Siguaw, J.A. (2000). *Introducing LISREL*. London: Sage Publications.

Grover, S., & Basu, S. (2017, March). Measuring student learning in introductory block-based programming: Examining misconceptions of loops, variables, and Boolean logic. In *Proceedings of the 2017 ACM SIGCSE Technical Symposium on Computer Science Education* (pp. 267-272). ACM.

Hu, L. T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, *6*(1), 1-55.

K–12 Computer Science Framework. (2016). *K–12 Computer Science Framework*. Retrieved from http://www.k12cs.org.

Raykov, T. (1997). Estimation of composite reliability for congeneric measures. *Applied Psychological Measurement*, *21*(2), 173-184.

Robins, A., Rountree, J., & Rountree, N. (2003). Learning and teaching programming: A review and discussion. *Computer Science Education*, *13*(2), 137-172.

Tabachnick, B.G. & Fidell, L.S. (2007). *Using Multivariate Statistics (5th ed.)*. Boston, MA: Allyn & Bacon/ Pearson Education.

Weintrop, D., & Wilensky, U. (2015, July). Using commutative assessments to compare conceptual understanding in blocks-based and text-based programs. In *ICER* (Vol. 15, pp. 101-110).

Wright, B. D., & Linacre, J. M. (1994). Reasonable mean-square fit values. *Rasch Measurement Transactions*, *8* (3), 370.