DOI: 10.33965/celda2019_201911L023
16th International Conference on Cognition and Exploratory Learning in Digital Age (CELDA 2019)

MEASURING COMPUTATIONAL THINKING – ADAPTING A PERFORMANCE TEST AND A SELF-ASSESSMENT INSTRUMENT FOR GERMAN-SPEAKING COUNTRIES

Josef Guggemos[1], Sabine Seufert[1] and Marcos Román-González[2]
[1]*University of St.Gallen, Guisanstr 1a, 9010 St.Gallen, Switzerland*
[2]*Universidad Nacional de Educación a Distancia, C/ Juan del Rosal 14, C.P. 28040, Madrid, Spain*

ABSTRACT

Far-reaching technological changes are shaping our society and the ways in which we work. A key 21st-century skill for taking advantage of those changes may be computational thinking (CT). CT aims at enabling humans to carry out more effective problem solving by utilizing concepts of computing and computer technology. For a successful integration of CT into curricula, however, it is important to take assessment into account. We review two instruments that capture CT: the Computational Thinkig Test (CTt), a performance test, and the Computational Thinking Scales (CTS), that relies on self-assessment. We have adapted both instruments from English to German. Using a sample of 202 upper-secondary students from Switzerland, we provide further evidence on the validity of both instruments. To this end, we apply item response theory and confirmatory factor analysis. Furthermore, we evaluate the relationship between CTt and CTS. Both instruments show good properties and may be suitable for assessing CT in German-speaking countries at the secondary level.

KEYWORDS

Computational Thinking, Rasch-Scaling, Performance Test, Self-Assessment

1. INTRODUCTION

Tremendous technological changes are shaping our society and ways of working (Weng, 2015). A key driver for these changes are advances in the field of computing (Wing, 2008). From an educational point of view, it is vital to determine skills that are necessary for being successful in such an environment. Among others, computational thinking (CT) is regarded as a key 21st-century skill (Voogt, Fisser, Good, Mishra, & Yadav, 2015; Wing, 2006).

Already used in the 1980s, the term CT experienced a revival in 2006 due to Wing's seminal article. Wing conceptualized CT as "solving problems, designing systems, and understanding human behavior, by drawing on the concepts fundamental to computer science" (2006, p. 33). Research activity on CT has experienced a sharp increase in recent years. Using Scopus as the database, Hsu et al. (2018) reviewed 120 CT studies published from 2006 to 2017. Of these studies, 65 were published in 2016 and 2017. None of the 120 articles relate to German-speaking countries. However, CT is also a construct of interest in such countries (for Germany, e.g., Delcker & Ifenthaler, 2017; for Switzerland, e.g., Repenning, 2018). Moreover, CT was part of the 2018 'International Computer and Information Literacy Study' (ICILS) in Germany (Eickelmann, 2019).

On behalf of the European Commission, Bocconi et al. (2016) reviewed the curricular integration of CT in European countries. Concerning the German-speaking countries, they provide the following evidence: In Austria, elements of CT are part of 'Informatics' in secondary schools. In the German-speaking part of Switzerland, the curricula for primary and lower-secondary schools comprise elements of CT. Furthermore, depending on the canton, CT may also be part of upper-secondary schools (e.g., Canton of St.Gallen, 2018). In Germany, curricula are managed on a regional level; hence, it is difficult to come to an overall conclusion. Examples of federal states where CT is part of the curriculum are Bavaria and North-Rhine Westphalia. On an abstract level, CT might be a curricular goal in German-speaking countries. However, as Grover and Pea (2013, p. 41) concluded: "Without attention to assessment, CT can have little hope of making its way

successfully into any K–12 curriculum." Recently, two literature reviews on CT stressed the importance of valid tools for measuring CT (Hsu et al., 2018; Shute, Sun, & Asbell-Clarke, 2017). Against this background, we raise the following research question: *How can CT of secondary students be assessed in German-speaking countries?*

We do not aim at large-scale assessment, like the ICILS study, but instead focus on instruments that researchers and educators can use to measure CT at the secondary level. Due to the need for CT interventional studies (Lye & Koh, 2014), instruments should be suitable for pre-post or longitudinal designs.

For a valid measurement, curriculum, instruction, and assessment have to be in alignment (Pellegrino, 2010). Hence, we will structure our review of the literature in line with the curriculum-instruction-assessment triad.

2. STATE OF THE ART

2.1 Curriculum

The basis for a curriculum may be a robust and agreed definition of CT (Selby & Woollard, 2013). Although such a definition is (and will be) not available, a core set of CT facets might be identified (Bocconi et al., 2016). Table 1 summarizes this core set. Since the core facets presented by Bocconi et al. (2016) are consistent with CT curricula in European countries, we expect a curricular-valid test to be in line with those facets.

Table 1. Core facets of CT

Abstraction	"simplifying from the concrete to the general as solutions are developed" (Barr & Stephenson, 2011, p. 52)
Algorithmic thinking	using "a step-by-step procedure for taking input and producing some desired output" (Wing, 2008, p. 3718)
Automation	"process in which a computer is instructed to execute a set of repetitive tasks quickly and efficiently compared to the processing power of a human" (Lee et al., 2011, p. 33)
Decomposition	"breaking problems down into smaller parts that may be more easily solved" (Barr & Stephenson, 2011, p. 52)
Debugging	"find your own mistakes and fix them" (Hsu et al., 2018, p. 299)
Generalization	"move from specific to broader applicability" (Selby & Woollard, 2013, p. 4)

2.2 Instruction

The crucial question regarding CT instruction is whether coding is necessary to foster CT. Buitrago Flórez et al. (2017) concluded that teaching programming might be a suitable approach for fostering CT. However, using professional programming languages like Java can be extremely difficult for secondary students due to complex syntax. It may be preferable to use visual programming languages (Lye & Koh, 2014; Repenning, 2017). Scratch, developed by the MIT Media lab (https://scratch.mit.edu), is such a visual programming language and is freely available in a German version. Scratch reduces cognitive load because it relies on visual blocks to build code and therefore prevents syntax errors. Students arrange the blocks via drag and drop.

In terms of instructional practice, Hsu et al. (2018) found that in 52% of all cases, a programming language is used to foster CT (other approaches are experiments and computer games). Scratch accounts for 41%, by far the highest share. The professional programming language with the highest share is MATLAB (4%). Overall, professional programming languages play only a minor role (15%).

2.3 Assessment

Román-González, Moreno-León, and Robles (2019) presented a classification for CT assessment tools. In line with this classification, two kind of assessment tools may be suitable for a pre-post design: Diagnostic tools address the "CT aptitudinal level"; perceptions-attitudes scales capture "the perceptions (e.g., self-efficacy perceptions) and attitudes of the subjects not only about CT, but also about related issues such as computers, computer science, computer programming, or even digital literacy" (Román-González et al., 2019, pp. 81-83).

Román-González et al. (2017, 2018a, 2018b) have developed and validated a *diagnostic tool* for secondary students: the Computational Thinking Test (CTt). Figure 1 shows a sample item. The authors rely on the CT framework of Brennan and Resnick (2012). It is consistent with the core CT facets. 'Abstraction' is covered as visual code blocks represent the problems, including conditionals and variables. 'Algorithmic thinking' is necessary because all tasks require sequencing steps to come to a solution. 'Automation' is captured by means of loops. 'Decomposition' manifests itself in the use of functions to split up the problems into more manageable elements. 'Debugging' is required because students have to identify mistakes in provided sequences of code blocks. 'Generalization', however, is not directly addressed by CTt. For assembling visual code blocks, the authors utilize the 'code.org' platform (https://code.org/), which is similar to Scratch (Hsu et al., 2018, p. 302). This might be favorable in terms of alignment with instruction. The test comprises 28 selected response items and can be taken online; secondary students should be able to process CTt in less than 45 minutes. No programming experience is necessary, which makes CTt a very flexible instrument. Since 'code.org' is licensed under Creative Commons, new items could be constructed with reasonable effort. The authors validated CTt using a sample of 1,251 Spanish secondary (5^{th} to 10^{th} grade) students and classical test theory. The reliability of the test is sufficiently high (Cronbach's alpha = .79). Concerning criterion validity, the authors investigated the nomological net of CT (Román-González et al., 2017; Román-González et al., 2018b). They report high correlations with problem-solving ability (r = .67) and moderate correlation with reasoning ability (r = .44), spatial ability (r = .44), and CT self-efficacy (r = .41). Moreover, there is evidence for instructional sensitivity of CTt (Brackmann et al., 2017; Rose, Habgood, & Jay, 2019). This means that responses of students to the items change due to instruction (Naumann, Rieser, Musow, Hochweber, & Hartig, 2019).

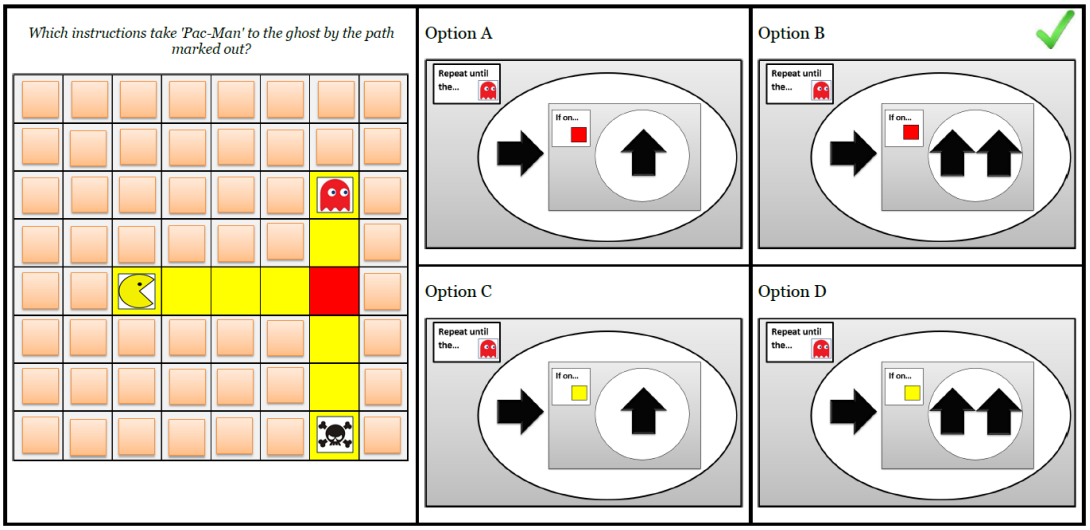

Figure 1. Item 6 of CTt

In sum, we regard CTt as a suitable instrument for measuring the CT of secondary students. As the authors acknowledge (Román-González et al., 2018b), the focus of CTt is narrow. However, it covers core CT facets and therefore might be a good trade-off between curricular validity, reliability, and required test time. Currently, CTt has only been validated using classical test theory. Using item response theory (IRT)

may allow further insights. In principal, CTt should be suitable for IRT methods because all items can be solved independently from each other; local stochastic independence of the items might be fulfilled.

As the focus of CTt is narrow, a *perceptions-attitudes scale* may complement CTt. The Computational Thinking Scales (CTS) are such an instrument (Korkmaz, Çakir, & Özden, 2017). The authors draw on the 'International Society for Technology in Education' (ISTE, 2015) framework of computational thinking that comprises five dimensions: 'Creativity', 'Algorithmic thinking', 'Cooperativity', 'Critical thinking', and 'Problem solving'. CTS consists of 29 self-assessment questions. It has been validated by means of confirmatory factor analysis using a sample of 580 Turkish undergraduate students. Fit-values are decent (Korkmaz et al., 2017, p. 565): CFI = 0.95, RMSEA = 0.06. Overall, we regard CTS as an appropriate instrument for capturing CT. Cost efficiency is high, because it requires only about five minutes test time.

3. METHOD

3.1 Adaption of CTt and CTS

Since CTt and CTS show promising results in validation studies and are suitable for capturing CT in longitudinal designs, we have adapted them to German. In terms of CTt, we utilized 'code.org' to create the visual code blocks in German. We also replaced the five easiest items by five more difficult items. This is possible because the original version of CTt comprises 40 items. An online version of CTt in German is available for Unipark from the authors upon request. A short version of CTS can be found in the appendix.

3.2 Psychometric Test Validation of CTt

We go beyond the work of the test developers and rely on probabilistic test theory. Due to the favorable characteristic of specific objectivity, we aim at Rasch modeling. The main advantage is that students (proficiency) and items (difficulty) can be located on a common (logit) scale that allows for a criterion-referenced test interpretation (Hartig & Frey, 2013). For assessing Rasch scalability of CTt, we draw on the framework of Bühner (2011, p. 547). First, we carry out Andersen's likelihood ratio test (LRT) (Andersen, 1973) using the R package 'eRm 0.16-2' (Mair & Hatzinger, 2007). In order to perform the LRT, the persons in the sample are split up into subgroups. Then, it is tested if the items as a whole work in the same way in these subgroups, i.e., the same parameter estimations are obtained. The median of the CTt raw score, gender, age (above and below average), and computer literacy (above and below average) act as split criteria. Computer literacy is measured with the dimension 'practical computer knowledge' of the INCOBI-R (Richter, Naumann, & Horz, 2010). Based on DIF-analyses, we may exclude items that discriminate specific groups.

To check for item homogeneity (unidimensionality), we conduct confirmatory factor analysis ('lavaan' 0.6-3 package in R, Rosseel, 2012) with CT as a single factor. Since the data are binary (correct/wrong), we use a 'WLSMV' estimator. A chi-square test acts as a global fit test. Furthermore, we rely on CFI, TLI, RMSEA, and SRMR as fit measures. Cutoff values for a decent fit may be: CFI and TLI > .95, RMSEA < .08, and SRMR < .11 (Bühner, 2011, pp. 425–427). Moreover, we analyze the correlation of the residuals by means of principal component analysis (PCA). This approach would also indicate violations of local stochastic independence of the items (OECD, 2017, pp. 169–170). For assessing person homogeneity, we perform a mixed Rasch analysis ('mixRasch' 1.1 package in R, Willse, 2011).

After having checked Rasch scalability, we examine if the items meet the cut-off values applied in the PISA studies (OECD, 2017, pp. 131–134) using the R package 'TAM 3.1-45' (Robitzsch, Kiefer, & Wu, 2019). The deviance from the item discrimination implied by the Rasch model is evaluated by means of weighted mean square error (wMNSQ = Infit). It should lie between 0.8 and 1.2. The point-biserial correlation should be above 0.30. The percentage of correct answers should fall between 20 and 90%. Not more than 10% of missing data should be present. Finally, we provide EAP/PV- and WLE-reliability as a measure for overall test reliability. These measures are comparable with Cronbach's alpha.

3.3 Psychometric Test Validation of CTS

CTS items are measured on a 7-point scale of rating. Hence, we use confirmatory factor analysis with a 'MLR' estimator for test validation ('lavaan' package 0.6-3 in R, Rosseel, 2012). A chi-square test acts as a global fit test. Furthermore, we rely on CFI, TLI, RMSEA, and SRMR as fit measures. Convergent validity is assessed by means of the average variance extracted (AVE). An AVE of greater than 0.5 may be evidence for convergent validity (Hair, Ringle, & Sarstedt, 2011, p. 145). Discriminant validity may be ensured if the square root of the AVE of every construct is higher than all correlations with other constructs (Fornell-Larcker criterion). To assess the reliability of the measures, we use composite reliability in conjunction with Cronbach's alpha. They should be higher than 0.7 (Hair et al., 2011, p. 147). Since the five facets are, as the name implies, facets of CT, we also checked if a second-order model (Bagozzi & Yi, 2012) with CT operationalized by the five facets yields a decent fit.

3.4 Sample

Two hundred and two upper-secondary students from German-speaking Switzerland act as a sample. They all attended the 11th (second last) grade at a 'Kantonsschule' (high school). On average, they were 17.23 years old (SD = 0.85 years) and 56% are female. CTt, CTS, and context questions were administered using Unipark at the beginning of the school year 2018/19. Teachers supervised the students and ensured an adequate test environment, e.g., preventing copying from their neighbor. The intended test time was 45 minutes. Ninety-five percent of the students were able to finish the instrument within this time; teachers allowed every student to complete the work.

4. RESULTS

4.1 Psychometric Validity of CTt

Of the 28 items, the students in the sample answered on average 18.45 items correctly (SD = 5.71, median = 19, min = 6, max = 28). Concerning Rasch-scalability, the LRT yielded mixed results. We did not find significant DIF-effects in terms of gender ($\chi^2 = 36$, df = 27, p = .11), age ($\chi^2 = 16$, df = 26, p = .94), and computer literacy ($\chi^2 = 30$, df = 26, p = .26). However, utilizing the median of the CTt test score as split criterion yielded significant DIF ($\chi^2 = 77$, df = 27, $p < .01$). Three items caused this overall DIF-effect. Item 1 was far too easy for the students in our sample (-4.80 Logits). Item 10 may have caused problems due to a different response format. The provided answer 'Option A and C are correct' might have confused students. For item 20, we could not find a reason on the content level. Moreover, the DIF-effect was only light to moderate (Penfield & Algina, 2006): Logit = 0.51. Against this background and because content validity was not impaired, we decided to exclude items 1 and 10 from the test and retain item 20. All further analysis was carried out without these two items.

The assumption of item homogeneity (unidimensionality) of CTt is justified: $\chi^2(199) = 341$ (p = .05), CFI = 0.964, TLI = 0.961, RMSEA = 0.026, SRMR = 0.063. Moreover, the PCA of the residual correlations yielded a percentage of variance for the first principal component of only 7%. This finding might also indicate local stochastic independence of the items. Our mixed Rasch analysis revealed a one-class solution; the AIC is lower in comparison to any multiclass solution (Bühner, 2011, p. 547). Hence, deviant problem-solving patterns among the students may be unlikely.

Concerning the cut-off values from the PISA studies, in general, all items show good values. The wMNSQ lies between 0.89 and 1.15 with the exception of item 18. This item has a wMNSQ of 1.22, which is slightly above the cut-off value of 1.2. However, wMNSQ values up to 1.33 might be acceptable (Wilson, 2005, p. 129). All point-biseral correlations are higher than 0.30. The percentage of correct answers for all items lies between .90 and .25. Every student fully processed the items; missing values are not present. Table 2 summarizes the item characteristics. EAP/PV-reliability equals 0.85, WLE-reliability 0.81, which may be sufficiently high for research purposes.

4.2 Psychometric Validity of CTS

In our sample, the fit-values of the published version of CTS with 29 questions indicated room for improvement ($\chi^2(340) = 657$ ($p < .001$), CFI = 0.881, TLI = 0.868, RMSEA = 0.073, SRMR = 0.095). The reasons are mainly cross loadings. For instance, the first item of 'Algorithmic thinking' also loads significantly on 'Critical thinking' and 'Creativity'. Discriminant validity is not ensured. Based on a content review, we selected three items for each of the five facets. This approach yielded a decent fit: $\chi^2(80) = 85$ ($p = .34$), CFI = 0.997, TLI = 0.996, RMSEA = 0.018, SRMR = 0.040. Convergent and discriminant validity are fulfilled: AVE is greater than .543 in every case, the Fornell-Larcker criterion is met. The five facets are reliably measured. The key characteristics of our short version of CTS can be found in Table 3. A second-order model with an overall CT factor yielded a 'Heywood case', i.e., the variance of 'critical thinking' is (non-significant) negative. Applying (meaningful) model restrictions did not solve this problem. Against this backdrop and overall small correlations between the five facets, e.g., between 'Algorithmic thinking' and 'Cooperative thinking', a second-order model might not be justified (Bagozzi & Yi, 2012).

Table 2. Item difficulty, fit, and DIF-effects of CTt (n = 202)

	Item difficulty and discrimination					DIF in Logit			
Item	θ	s.e. θ	wMNSQ	Pt.bis.	P+	Ability	Gender	Age	CL
ct_2	-2.42	0.23	0.99	0.34	87%	-0.04	-0.09	-0.01	-0.13
ct_3	-2.75	0.25	0.99	0.30	90%	-0.06	0.41	0.18	-0.30
ct_4	-0.32	0.16	1.14	0.42	55%	-0.37	0.18	-0.17	0.07
ct_5	-0.45	0.16	1.08	0.45	58%	-0.20	-0.13	0.03	-0.04
ct_6	-2.22	0.21	1.07	0.33	85%	-0.07	0.17	0.09	-0.09
ct_7	-0.18	0.16	0.97	0.54	53%	0.23	0.15	0.21	-0.25
ct_8	-1.24	0.18	1.15	0.34	72%	-0.45	-0.37	-0.01	-0.06
ct_9	-1.63	0.19	0.97	0.44	78%	-0.11	0.10	0.26	-0.41
ct_11	-0.72	0.17	0.92	0.56	63%	0.33	0.43	0.02	-0.06
ct_12	-0.80	0.17	1.09	0.41	64%	-0.23	-0.22	-0.10	-0.07
ct_13	-1.40	0.18	1.05	0.39	74%	-0.25	-0.37	0.41	-0.25
ct_14	-0.94	0.17	1.07	0.42	67%	-0.23	-0.03	0.03	0.17
ct_15	-2.01	0.20	0.91	0.45	83%	0.38	0.10	-0.04	0.32
ct_16	-1.53	0.18	0.98	0.46	76%	0.05	0.35	-0.33	0.21
ct_17	-0.18	0.16	0.94	0.57	53%	0.05	0.26	-0.16	0.18
ct_18	1.45	0.19	1.22	0.30	25%	-0.58	-0.13	0.33	0.08
ct_19	-1.85	0.20	0.92	0.47	81%	0.16	0.03	-0.03	0.16
ct_20	-0.69	0.17	0.89	0.58	62%	0.51	0.32	-0.06	0.02
ct_21	-0.94	0.17	0.94	0.53	67%	0.38	-0.25	-0.11	0.26
ct_22	-0.91	0.17	1.04	0.46	66%	0.04	-0.17	0.04	0.00
ct_23	-1.40	0.18	1.01	0.45	74%	-0.18	-0.31	-0.18	0.02
ct_24	-0.50	0.16	0.89	0.59	59%	0.14	-0.12	-0.19	0.33
ct_25	-1.46	0.18	0.91	0.52	75%	0.29	-0.29	-0.18	0.08
ct_26	-0.26	0.16	0.93	0.57	54%	0.20	-0.03	0.12	-0.17
ct_27	0.51	0.17	0.96	0.56	40%	0.00	0.05	-0.12	0.09
ct_28	-0.10	0.16	0.97	0.55	51%	0.03	-0.01	-0.02	-0.16

Note. θ = difficulty, s.e. = standard error, wMNSQ = weighted mean square error, Pt.bis. = point biserial correlation, P+ = correct responses, Ability = median of CTt raw score, CL = computer literacy (INCOBI-R).

4.3 Comparison CTt and CTS

The latent correlations between CTt and the five facets of CTS are with 'Creativity' .271 ($p = .002$), 'Algorithmic thinking' .309 ($p < .001$), 'Cooperativity' -.003 ($p = .956$), 'Critical thinking' .408 ($p < .001$), and 'Problem solving' .154 ($p = .085$). Considering all CTS facets (latent scores) as independent variables and CTt as a dependent variable in a latent regression, only algorithmic thinking is statistically significant ($b = 0.319$, $p < .001$).

Table 3. Characteristics of used CTS items and constructs (n = 202)

	Construct	Item	Mean (SD)	λ	α	ρc	AVE	Latent correlations, square root of AVE on diagonal				
								(1)	(2)	(3)	(4)	(5)
(1)	Creativity	cr_3	5.6 (1.5)	.75								
		cr_4	5.6 (1.3)	.86	.87	.94	.75	.87				
		cr_5	5.3 (1.3)	.89								
(2)	Algorithmic	al_3	4.0 (1.9)	.90								
	thinking	al_4	3.7 (1.8)	.86	.90	.94	.73	**.32**	.85			
		al_5	4.0 (1.8)	.83								
(3)	Cooperativity	co_1	4.6 (1.8)	.85								
		co_2	4.2 (1.8)	.90	.89	.93	.70	**.22**	.15	.84		
		co_3	4.8 (1.7)	.82								
(4)	Critical	cr_1	5.1 (1.4)	.74								
	thinking	cr_2	4.9 (1.4)	.75	.80	.88	.57	**.71**	**.62**	**.22**	.75	
		cr_3	4.6 (1.4)	.78								
(5)	Problem	pr_1	5.5 (1.5)	.79								
	solving	pr_2	5.4 (1.7)	.68	.78	.86	.54	.20	-.03	-.20	.15	.73
		pr_4	5.1 (1.5)	.74								

Note. Items measured on a 7-point rating scale. λ = standardized loading, α = Cronbach's alpha, ρc = composite reliability, AVE = average variance extracted. Figures in bold indicate significant correlations at the 5% level.

5. CONCLUSION

Recent literature reviews have stressed the theoretical and practical importance of CT. Moreover, CT is a curricular goal in German-speaking countries. The current shortage of quantitative research on CT in German speaking countries may be attributed to the lack of instruments for measuring CT. To tackle this issue, we have adapted two internationally accepted (Shute et al., 2017) instruments for assessing CT and evaluated them using a sample of 202 upper-secondary students from German-speaking Switzerland. CTt, as a diagnostic tool, covers core CT facets addressed in curricula. Against this background, curricular validity might be ensured. Secondary students should be able to perform CTt in less than 45 minutes. CTt relies on dichotomous constructed response items. This allows an objective and very cost-efficient scoring. On the downside, CTt may not be able to capture higher-level cognitive processes.

In terms of psychometric validity, we were able to demonstrate Rasch scalability. The overall EAP/PV- and WLE-reliability of the test is good with values of 0.85 and 0.81, respectively. In sum, CTt may be adequate for summative CT assessment of secondary students in German-speaking countries, e.g., for research purposes, especially because it relies on visual code blocks that most of the applied instructional means also utilize. Since instructional sensitivity of CTt has been demonstrated in other studies, it may be appropriate for measuring CT changes in (quasi-)experimental designs.

In contrast to CTt, the perceptions-attitudes scale CTS covers CT on a broad scale as it relies on the ISTE (2015) framework. 'Creativity' and 'Cooperativity', two of the five facets of CTS, might be constructs related to, but not parts of, CT. The overall low and even negative latent correlations among the five facets, and the convergence problems in case of a second-order model, might be evidence for this assertion. Furthermore, the original version of CTS showed a lack of discriminant validity. We addressed this issue by removing items while considering content validity. In this light, we may have contributed to the psychometric validity of CTS. Overall, we regard CTS as a suitable approach for capturing CT in a longitudinal design. The main advantage is the small amount of time necessary for our short version, with only 15 questions. Students should be able to perform it in less than three minutes. This makes CTS also an attractive option for considering CT as a control variable. The main disadvantage of CTS lies in the nature of self-assessments – it is questionable whether students are able and willing to evaluate themselves accurately. Overall, we agree with Román-González et al. (2019) that only a combination of instruments may yield a comprehensive picture of CT. Since CTt and CTS both are suitable for longitudinal studies, they might be used for capturing CT in a pre-post design in German speaking countries.

REFERENCES

Andersen, E. B. (1973). A goodness of fit test for the Rasch model. *Psychometrika*, *38*(1), 123–140.

Bagozzi, R. P. & Yi, Y. (2012). Specification, evaluation, and interpretation of structural equation models. *Journal of the Academy of Marketing Science*, *40*(1), 8–34.

Barr, V., & Stephenson, C. (2011). Bringing computational thinking to K-12. *ACM Inroads*, *2*(1), 48–54.

Bocconi, S., Chioccariello, A., Dettori, G., Ferrari, A., & Engelhardt, K. (2016). *Developing computational thinking in compulsory education*. Luxembourg: Publications Office of the European Union.

Brackmann, C. P., Román-González, M., Robles, G., Moreno-León, J., Casali, A., & Barone, D. (2017). Development of Computational Thinking Skills through Unplugged Activities in Primary School. In E. Barendsen (Ed.), *Proceedings of the 12th Workshop on Primary and Secondary Computing Education* (pp. 65–72). New York, NY: ACM.

Brennan, K., & Resnick, M. (2012). New frameworks for studying and assessing the development of computational thinking. *American Educational Research Association Meeting, Vancouver, BC, Canada*, 1–25.

Bühner, M. (2011). *Einführung in die Test- und Fragebogenkonstruktion* (3rd ed.). Munich: Pearson Studium.

Buitrago Flórez, F., Casallas, R., Hernández, M., Reyes, A., Restrepo, S., & Danies, G. (2017). Changing a Generation's Way of Thinking: Teaching Computational Thinking Through Programming. *Review of Educational Research*, *87*(4), 834–860.

Canton of St.Gallen. (2018). *Lehrplan Informatik*. St. Gallen: Amt für Mittelschulen.

Delcker, J. & Ifenthaler, D. (2017). Computational Thinking as an Interdisciplinary Approach to Computer Science School Curricula: A German Perspective. In P. J. Rich & C. B. Hodges (Eds.), *Emerging Research, Practice, and Policy on Computational Thinking* (pp. 49–62). Cham: Springer International Publishing.

Eickelmann, B. (2019). Measuring Secondary School Students' Competence in Computational Thinking in ICILS 2018—Challenges, Concepts, and Potential Implications for School Systems Around the World. In S.-C. Kong & H. Abelson (Eds.), *Computational thinking education* (pp. 53–64). Singapore: Springer.

Grover, S., & Pea, R. (2013). Computational thinking in K-12: A review of the state of the field. *Educational Researcher*, *42*(1), 38–43.

Hair, J. F., Ringle, C. M., & Sarstedt, M. (2011). PLS-SEM: Indeed a silver bullet. *The Journal of Marketing Theory and Practice*, *19*(2), 139–152.

Hartig, J., & Frey, A. (2013). Sind Modelle der Item-Response-Theorie (IRT) das „Mittel der Wahl" für die Modellierung von Kompetenzen? *Zeitschrift für Erziehungswissenschaft*, *16*, 47–51.

Hsu, T.-C., Chang, S.-C., & Hung, Y.-T. (2018). How to learn and how to teach computational thinking: Suggestions based on a review of the literature. *Computers & Education*, *126*, 296–310.

ISTE. (2015). Computational thinking: leadership toolkit. Retrieved from www.iste.org/computational-thinking

Korkmaz, Ö., Çakir, R., & Özden, M. Y. (2017). A validity and reliability study of the computational thinking scales (CTS). *Computers in Human Behavior*, *72*, 558–569.

Lee, I., Martin, F., Denner, J., Coulter, B., Allan, W., Erickson, J., . . . Werner, L. (2011). Computational thinking for youth in practice. *ACM Inroads*, *2*(1), 32.

Lye, S. Y. & Koh, J. H. L. (2014). Review on teaching and learning of computational thinking through programming: What is next for K-12? *Computers in Human Behavior*, *41*, 51–61.

Mair, P., & Hatzinger, R. (2007). Extended Rasch modeling: The eRm package for the application of IRT models in R. *Journal of Statistical Software*, *20*(9), 1–20.

Naumann, A., Rieser, S., Musow, S., Hochweber, J., & Hartig, J. (2019). Sensitivity of test items to teaching quality. *Learning and Instruction*, *60*, 41–53.

OECD. (2017). *PISA 2015 Technical Report*. Paris: OECD Publishing. Retrieved from http://www.oecd.org/pisa/sitedocument/PISA-2015-technical-report-final.pdf

Pellegrino, J. W. (2010). *The design of an assessment system for the race to the top: A learning sciences perspective on issues of growth and measurement*. Princeton: Educational Testing Service.

Penfield, R. D. & Algina, J. (2006). A generalized DIF effect variance estimator for measuring unsigned differential test functioning in mixed format tests. *Journal of Educational Measurement*, *43*(4), 295–312.

Repenning, A. (2017). Moving beyond syntax: Lessons from 20 years of blocks programing in AgentSheets. *Journal of Visual Languages and Sentient Systems*, *3*(1), 68–91.

Repenning, A. (2018). Scale or fail: Moving beyond self-selected computer science education in Switzerland. *Communications of the ACM*, *61*(5), 40–42.

Richter, T., Naumann, J., & Horz, H. (2010). Eine revidierte Fassung des Inventars zur Computerbildung (INCOBI-R). *Zeitschrift Für Pädagogische Psychologie*, *24*(1), 23–37.

Robitzsch, A., Kiefer, T., & Wu, M. (2019). Package 'TAM'. https://cran.r-project.org/web/packages/TAM/ TAM.pdf

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). Singapore: Springer.

Román-González, M., Pérez-González, J.-C., & Jiménez-Fernández, C. (2017). Which cognitive abilities underlie computational thinking? Criterion validity of the Computational Thinking Test. *Computers in Human Behavior*, *72*, 678–691.

Román-González, M., Pérez-González, J.-C., Moreno-León, J., & Robles, G. (2018a). Can computational talent be detected? Predictive validity of the Computational Thinking Test. *International Journal of Child-Computer Interaction*, *18*, 47–58.

Román-González, M., Pérez-González, J.-C., Moreno-León, J., & Robles, G. (2018b). Extending the nomological network of computational thinking with non-cognitive factors. *Computers in Human Behavior*, *80*, 441–459.

Rose, S. P., Habgood, M. J., & Jay, T. (2019). Using Pirate Plunder to Develop Children's Abstraction Skills in Scratch. In R. Mandryk, M. Hancock, M. Perry, & A. Cox (Eds.), *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems* (pp. 1–6). New York, New York, USA: ACM Press.

Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software*, *48*(2), 1–36.

Selby, C. C. & Woollard, J. (2013). *Computational thinking: The developing definition*. University of Southampton.

Shute, V. J., Sun, C., & Asbell-Clarke, J. (2017). Demystifying computational thinking. *Educational Research Review*, *22*, 142–158.

Voogt, J., Fisser, P., Good, J., Mishra, P., & Yadav, A. (2015). Computational thinking in compulsory education: Towards an agenda for research and practice. *Education and Information Technologies*, *20*(4), 715–728.

Weng, W. (2015). Eight skills in future work. *Education*, *135*(4), 419–422.

Willse, J. T. (2011). Mixture Rasch models with joint maximum likelihood estimation. *Educational and Psychological Measurement*, *71*(1), 5–19.

Wilson, M. (2005). *Constructing measures: An item response modeling approach*. New York: Psychology Press.

Wing, J. M. (2006). Computational thinking. *Communications of the ACM*, *49*(3), 33–35.

Wing, J. M. (2008). Computational thinking and thinking about computing. *Philosophical Transactions. Series A, Mathematical, Physical, and Engineering Sciences*, *366*(1881), 3717–3725.

APPENDIX

Short version of CTS (Korkmaz et al., 2017)

Creativity	"I believe that I can solve most of the problems I face if I have sufficient amount of time and if I show effort."
	"I have a belief that I can solve the problems possible to occur when I encounter with a new situation."
	"I trust that I can apply the plan while making it to solve a problem of mine."
Algorithmic thinking	"I think that I learn better the instructions made with the help of mathematical symbols and concepts."
	"I can mathematically express the solution ways of the problems I face in the daily life."
	"I can digitize a mathematical problem expressed verbally."
Cooperativity	"I like experiencing cooperative learning together with my group friends."
	"In the cooperative learning, I think that I attain/will attain more successful results because I am working in a group."
	"I like solving problems related to group project together with my friends in cooperative learning."
Critical thinking	"I am willing to learn challenging things."
	"I am proud of being able to think with a great precision."
	"I make use of a systematic method while comparing the options at my hand and while reaching a decision."
Problem solving	"I have problems in the demonstration of the solution of a problem in my mind." (R)
	"I have problems in the issue of where and how I should use the variables such as X and Y in the solution of a problem." (R)
	"I cannot apply the solution ways I plan respectively and gradually." (R)

Note. Selection of 15 out of 29 items (Korkmaz et al., 2017, p. 565). Measured on a 7-point rating-scale ranging from 'not true at all' to 'entirely true'. R = reverse coding.