



Computational Thinking Assessment – Towards More Vivid Interpretations

Josef Guggemos¹ · Sabine Seufert¹ · Marcos Román-González²

Accepted: 10 December 2021 / Published online: 27 January 2022
© The Author(s), under exclusive licence to Springer Nature B.V. 2021

Abstract

Computational thinking (CT) is an important 21st-century skill. This paper aims at more useful CT assessment. Available evaluation instruments are reviewed; two generally accepted CT evaluation tools are selected for a comprehensive CT assessment: the CTt, a performance test, and the CTS, a self-assessment instrument. The sample comprises 202 high school students from German-speaking Switzerland. Concerning the CTt, Rasch-scalability is demonstrated. Utilizing the approach of the PISA studies, proficiency levels are formed that comprise tasks with specific characteristics that students are systematically able to master. This could help teachers to offer individual support to their students. In terms of the CTS, the original version is refined using confirmatory factor and measurement-invariance analysis. A latent profile analysis yielded four profiles, two of which are of particular interest. One profile comprises students with, on the one hand, moderate to high creative thinking ability, cooperativity, and critical thinking skills and, on the other hand, low algorithmic thinking ability. The second remarkable profile consists of students with particularly low cooperativity. Based on these strength and weakness profiles, teachers could offer support tailored to student needs.

Keywords Computational thinking · Performance test · Item response theory · Latent profile analysis · Person-centered assessment · Proficiency level model

Abbreviations

aBIC	Adjusted Bayesian information criterion
AIC	Akaike's information criterion
BIC	Bayesian information criterion
BLRT	Bootstrap likelihood ratio test
CAIC	Consistent Akaike's information criterion
CFI	Comparative fit index
CI	Confidence interval

✉ Josef Guggemos
josef.guggemos@unisg.ch

¹ TechnologiesSchool of Management, Institute forEducational Management and Technologies, University of St.Gallen, St. Jakob-Strasse 21, 9000 St. Gallen, Switzerland

² Universidad Nacional de Educación a Distancia (UNED), Faculty of Education, C/ Juan del Rosal 14, C.P. 28040 Madrid, Spain

CT	Computational thinking
CTS	Computational Thinking Scale
CTt	Computational Thinking Test
DIF	Differential item functioning
EAP/PV	Expected a posteriori / plausible values
ICILS	International Computer and Information Literacy Study
LPA	Latent profile analysis
LRT	Andersen's likelihood ratio test
MANOVA	Multivariate analysis of variance
MLR	Maximum likelihood with robust standard errors
PISA	Programme for International Student Assessment
R ²	Coefficient of determination
RMSEA	Root mean square error of approximation
RQ	Research question
SRMR	Standardized root mean square residual
TLI	Tucker-Lewis Index
WLE	Weighted likelihood estimate
WLSMV	Diagonally weighted least squares
wMNSQ	Weighted mean square
YB- χ^2	Yuan-Bentler corrected χ^2
α	Cronbach's alpha
ω	Revell's omega total

1 Introduction

Tremendous technological changes are shaping our society and ways of working (Harteis et al., 2020; Ifenthaler et al., 2021; Kirschner & Stoyanov, 2020). In this context, computational thinking (CT) is regarded as a key 21st-century skill (Voogt et al., 2015; Wing, 2006; Yadav et al., 2016). The significance of CT in the twenty-first century may be evident (Barr et al., 2011; Buitrago Flórez et al., 2017; Wing, 2008): computing and computer technology pervade every field of study and workplace. CT aims at enabling humans to use these resources for solving problems.

In her seminal paper, Wing conceptualizes CT as “solving problems, designing systems, and understanding human behavior, by drawing on the concepts fundamental to computer science” (Wing, 2006, p. 33). As the theoretical basis for CT, the framework of Brennan and Resnick (2012) is often utilized. It comprises three dimensions (Brennan & Resnick, 2012, pp. 3–11). *Computational concepts* are common in many programming languages, but not restricted to them: sequences, loops, events, parallelism, conditionals, operators, and data. *Computational practices* are the processes in which students engage: imaging and building, testing and debugging, reusing and remixing, abstracting, and modularizing. *Computational perspectives* are shifts in points of view, relationships to others, and the digital world around: expressing, connecting, and questioning.

In CT research, the relationship between CT and programming is often thematized. Israel et al. (2015) regard the use of computers to model ideas and programming as an integral part of CT. Buitrago Flórez et al. (2017), as well as Lye and Koh (2014), argue that by means of programming, several core facets of CT can be addressed. Shute et al. (2017) concluded that there is a close relationship between CT and programming skills due to

similar underlying cognitive processes. Hsu et al. (2018), based on their review of the literature, reported that programming is widely used to teach CT. Grover et al. (2016) maintained that programming has a positive influence on the experience on CT. Scherer et al. (2019), based on a meta-analysis, concluded that CT can be taught through programming. However, using professional programming languages like Java can be extremely difficult for students due to complex syntax, and it may be preferable to use visual programming languages (Lye & Koh, 2014; Repenning, 2017). Scratch, developed by the Massachusetts Institute of Technology Media Lab (<https://scratch.mit.edu>), is such a visual programming language that is heavily used as an instructional tool (Hsu et al., 2018).

The importance of CT assessment is regularly stressed (Grover & Pea, 2013; Ilic et al., 2018; Shute et al., 2017; Tang et al., 2020; Weintrop et al., 2021). However, it has to be highlighted that assessment is not an end in itself, but it should contribute to promoting student learning (Pellegrino et al., 2016). When assessing complex skills such as CT, the structure as well as the levels of the construct have to be considered (Seufert et al., 2021). For instance, the DigComp 2.1 framework that addresses digital competencies comprises five dimensions (structure), e.g., information literacy, as well as eight proficiency levels ranging from foundation to highly specialized (Carretero et al., 2017). In terms of CT, research about proficiency levels is in its infancy. The importance of modeling proficiency levels for a better understanding of this construct, however, is stressed by the 2018 International Computer and Information Literacy Study (ICILS) (Fraillon et al., 2019). Without proficiency levels, test results are difficult to comprehend and to communicate. Findings such as ‘the student belongs to the top 10% of all test takers’ or ‘the student answered 67% of the questions correctly’ are not very helpful for the purpose of fostering student learning. A proficiency level model could allow for a more vivid interpretation of the test results by relying on items with clearly specified characteristics that students are systematically able to master (AERA et al., 2014). This also makes it possible to set operationalized learning goals. For example, the goal could be that all students in a class are systematically able to use functions. Students who do not reach the corresponding proficiency level could receive specific guidance and support.

Due to the multifaceted nature of CT, it is unlikely that a single instrument is sufficient to comprehensively capture CT (Polat et al., 2021; Román-González et al., 2019). Rather, a system of various assessments may be necessary. Using multidimensional approaches could also reveal different CT profiles. For instance, some students may perform well in the realm of computational concepts, but at the same time poorly in the area of computational perspectives. By solely focusing on a specific facet of CT, students with specific strength and weakness profiles may not be identified. Knowledge about CT profiles, i.e., different types of computational thinkers, however, could be utilized for personalized guidance and support (Hofmans et al., 2020). Although such person-centered assessments are widespread in other research fields (e.g., Lohr et al., 2021; Meyer & Morin, 2016; Scherer et al., 2021; Tondeur et al., 2019), they are not common in CT research. The benefits of such techniques, however, are acknowledged in CT research (Román-González et al., 2019).

2 Theoretical Perspectives

2.1 CT Assessment Instruments

Several authors have reviewed CT assessment instruments (Kong, 2019; Román-González et al., 2019; Shute et al., 2017; Tang et al., 2020). Román-González et al. (2019) developed a useful classification of assessment instruments (Israel-Fishelson & Hershkovitz, 2022). Following this classification, diagnostic tools that aim at capturing students' CT proficiency could be the most suitable basis for forming proficiency levels. Diagnostic tools are performance tests and do not require specific prior knowledge, e.g., a specific programming language. Hence, they can be used to evaluate learning gains by comparing pre- and post-test results, i.e., if students have reached a higher proficiency level after instruction.

A drawback of many diagnostic tools is that they are not freely available (ICILS 2018: Fraillon et al., 2019; Fairy Assessment: Werner et al., 2012; Basic Programming Abilities: Mühling et al., 2015). Furthermore, several tools focus on the pre-secondary level (Chen et al., 2017; Kong & Wang, 2021; Relkin et al., 2021; Seiter & Foreman, 2013); therefore, the covered proficiency spectrum is limited. From the freely available diagnostic tools, the Computational Thinking Test (CTt) may be especially suitable for the purpose of forming proficiency levels. The CTt (Román-González, 2015) is a performance test for secondary students using the framework of Brennan and Resnick (2012) as a theoretical background. It defines CT as “the ability to formulate and solve problems by relying on the fundamental concepts of computing, and using logic-syntax of programming languages: basic sequences, loops, iteration, conditionals, functions and variables” (Román-González et al., 2017, p. 681). Sample items can be found in Figs. 2 and 3. Due to the use of a visual programming language, the CTt covers a broad range of instructional settings. To perform the CTt no knowledge in a specific programming language, e.g., Java, is necessary, which makes it a very flexible instrument. The CTt comprises 28 selected response items and can be taken online; the target group should be able to process the test in less than 45 min. The CTt is claimed to be unidimensional although different cognitive operations are involved when performing the items. This is based on the notion of Fischer (1973) that the items of a unidimensional construct may be linearly decomposed into cognitive operations; this may also be the case for CT (Mühling et al., 2015). According to the definition of the CTt, cognitive operations could be *sequences*, *loops*, *conditionals*, *functions*, and *variables*. These correspond with the computational concepts dimension of the Brennan and Resnick (2012) framework.

Román-González et al., (2017) validated the CTt using a sample of 1,251 Spanish secondary students (5th to 10th grade) and classical test theory. The reliability of the test is sufficiently high (Cronbach's $\alpha=0.79$). Chan et al. (2020) provided evidence for Rasch scalability of the CTt based on a sample of 153 upper-secondary students from Singapore. The CTt is increasingly used in research projects for assessing CT learning (e.g., Guggemos, 2021; Brackmann et al., 2017; Hooshyar et al., 2021; Rose et al., 2019; Zhao & Shute, 2019). In light of this, relying on the CTt may be in line with the call for using standardized instruments to ensure comparability across studies (Shute et al., 2017).

Despite its suitability for assessing computational concepts and, to some degree, computational practices, the CTt also has disadvantages: it neglects computational perspectives (Román-González et al., 2017). For capturing such perspectives, perception-attitude scales may be suitable (Román-González et al., 2019). They capture self-efficacy beliefs by means of self-assessment. In general, complementing a performance test with a self-assessment

instrument may be beneficial to obtain a comprehensive picture of a construct (Rosman et al., 2015). Again, we aim at using standardized instruments and those with a specific focus on CT. This excludes computer attitude scales, e.g., Denner et al. (2014), Ericson and McKlin (2012), and Yadav et al. (2014). For the same reason, we do not consider generic self-efficacy or motivation scales. A viable option to capture computational perspectives may be the Computational Thinking Scales (CTS) (Polat et al., 2021). They were developed by Korkmaz et al. (2017) and is a standardized self-assessment instrument for capturing CT (Durak & Saritepeci, 2018; Israel-Fishelson & Hershkovitz, 2022; Shute et al., 2017). The authors utilize the International Society for Technology in Education (ISTE, 2015) framework of CT, namely, the five dimensions of *creativity*, *algorithmic thinking*, *cooperativity*, *critical thinking*, and *problem solving*. Descriptions of these dimensions can be found in Table 1.

The 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, the CTt and the CTS may be complementary assessment tools that can provide a comprehensive picture of students' CT ability (Polat et al., 2021). The CTt, as a unidimensional performance test, seems to be suitable to form proficiency levels, especially because the cognitive operations are documented. The CTS, as a multidimensional self-assessment instrument, can be used to identify CT profiles.

2.2 The Present Study

The current study aims at reaching a better understanding of CT as a construct. To this end, we contribute to more useful interpretations of CT assessment findings. Concerning performance tests, the importance of proficiency levels is stressed (Fraillon et al., 2019; OECD, 2017). However, the 2018 ICILS refrained from developing a proficiency level model due to the small number of CT test items in the study (Fraillon et al., 2020); only proficiency regions were described. When forming proficiency levels, referring to the cognitive operations involved when performing the items could be advantageous from a construct validity point of view (Embretson & Daniel, 2008). For instance, it would not be meaningful to form CT proficiency levels based on the text complexity of the items. In this regard, we can take advantage of the fact that for the CTt, the cognitive operations necessary for performing the items are documented. As the cognitive operations correspond with the computational practices of the Brennan and Resnick (2012) framework, this could be a sound theoretical basis. The paper at hand demonstrates that the cognitive operations can predict the difficulty of CT test items and can be utilized to form meaningful proficiency levels (Hartig et al., 2012). Our first research question is:

RQ1 What CT proficiency levels can be identified based on the CTt?

As already highlighted, the purpose of an assessment is to facilitate student learning. In the case of CT as a multifaceted construct, it is unlikely to observe only one CT profile. Rather, it can be expected to identify different types of computational thinkers. For teachers, knowledge about CT profiles could be helpful in reducing complexity. Students in the same profile could benefit from similar treatment (Hofmans et al., 2020); teachers may design their instructional measures around the identified profiles. Despite their usefulness for a better understanding of CT, to our knowledge, such person-centered methods of assessment have not yet been used in CT research. The CTS may be suitable for identifying latent profiles as it comprises five dimensions. Our second research question is:

RQ2 What CT profiles can be identified based on the CTS?

The CTt and CTS are regarded as complementary instruments that can offer a comprehensive picture of student CT (Polat et al., 2021; Román-González et al., 2019). Against this backdrop, we consider a third research question:

RQ3 How are the CTt and CTS results related to each other?

Since, recently, Polat et al. (2021) addressed the same research question, we will compare their findings with those of the present study.

3 Method

3.1 Sample and Adaption of Instruments

Our sample comprises 202 upper-secondary students from German-speaking Switzerland. They all attended the 11th (second last) grade at a *Kantonsschule* (high school), which is the most demanding school type in Switzerland. Data were collected at the beginning of the school year 2018/19. The CTt, CTS, and context questions were administered using Unipark. Teachers supervised the students and ensured an adequate test environment, e.g., preventing copying from their neighbor. The intended test time was 90 min. Ninety-five percent of the students were able to finish the test within this time; teachers allowed every student to complete the work. On average, the students were 17.23 years old ($SD=0.85$ years) and 56% were female. They experienced, on average, 2.89 h ($SD=1.20$ h) of computer science instruction in the past; students reported tigerjyton (<https://www.tigerjython.ch/engl/>) as the most generally used learning environment. Tigerjyton addresses important computational concepts such as sequencing, conditionals, functions, and loops. Overall, 77% of the students claim to be able to program, e.g., in Java or Python. To evaluate test motivation, we can draw information from two items within the context questionnaire: ‘When performing the tasks, I disengaged’ and ‘My mind was elsewhere when I was performing the tasks’ (Prenzel et al., 1998). Cronbach’s alpha equals 0.78. On average, the students disagreed concerning a lack in test motivation: $M=3.36$ and $SD=1.66$, based on a seven-point scale of rating. This is consistent with an absence of missing data; the omission of items can act as a proxy for a lack of test motivation (Ullrich et al., 2020). We also checked for multivariate outliers using Mahalanobis distances (Leys et al., 2019); no student is classified as an outlier at the 1% (and 5%) significance level.

Since the CTt was designed for 5th to 10th grade students, we replaced the five easiest items by five equivalent but more difficult ones. Equivalent means comparable in the environmental interface (canvas vs. maze), answer style, and required task. To this end, we drew on the initial pool of CTt items, namely forty, as well as where experts gauged an item difficulty (Román-González, 2015). The replacement of very easy items may be advantageous as they do not have an evidentiary value (Köhler & Hartig, 2017). A pre-test indicated that almost all the students within the target group (11th grade) would master the easiest five items. Figure 4 depicts an item that was integrated within our German CTt version in comparison to the original version. The item numbering in this paper always refers to our German version; the numbering allocation of the original CTt version can be

found in Table 2. To adapt the CTS from English to German we applied a back-translation approach (Maneesriwongul & Dixon, 2004).¹

3.2 RQ1: Forming Proficiency Levels

3.2.1 Psychometric Test Validation

Before forming proficiency levels, a psychometric test validation is necessary. If the CTt was Rasch scalable, this would imply specific objectivity: students (proficiency) and items (difficulty) can be located on a common Logit scale. This allows for a criterion-referenced test interpretation (Hartig & Frey, 2013). If the proficiency of a person equals the difficulty of an item (same location on the Logit scale), the expected probability of a correct response will be 50%. The proficiency of students can be described by referring to items that they are expected to master with a specified probability.² Specific objectivity would be violated, for instance, if some items were more difficult for males than for females.

For assessing Rasch scalability, 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 1.0–1' (Mair & Hatzinger, 2007). A significant LRT would indicate that the items work differently in specific subgroups, i.e., different parameter estimations for the difficulty of the items are obtained. In order to perform the LRT, the students in the sample have to be split up into subgroups. We use the median of the CTt raw score, gender (male vs. female), age (above and below average), computer literacy (above and below average), and ability to program (yes vs. no) as split criteria (Chan et al., 2020; Guggemos et al., 2019). Computer literacy is captured by the dimension *practical computer knowledge* of the INCOBI-R (Richter et al., 2010), and ability to program via student self-reporting. In case of a significant LRT, the next step is to check which items work differently in the subgroups. For instance, students who are able to program may have a systematic advantage in answering specific items (DIF-effect). Based on DIF-analyses ('TAM 3.5–19' package in R; Robitzsch et al., 2020), we may exclude items that systematically advantage or disadvantage specific subgroups. Test fairness (absence of DIF) is an important characteristic of an assessment instrument (AERA et al., 2014). In our case, it is of specific importance as it is the prerequisite for locating all students on one logit scale. In line with Penfield and Algina (2006, pp. 307–308), a DIF of less than 0.43 Logit may be negligible, between 0.43 and 0.64 Logit moderate, and above 0.64 Logit large.

The LRT and DIF-analyses rely on pre-specified split criteria. However, there may also be latent subpopulations of individuals for which the CTt works differently or who show deviant response behavior. Such latent subpopulations can be identified with a mixed Rasch analysis ('mixRasch' 1.1 package in R; Willse, 2011). An example of a latent subgroup could be students who are guessing in order to solve the selected response CTt items. If the mixed Rasch analysis reveals a one-class solution, this would be evidence for the overall fairness of the CTt. To identify the optimal number of latent classes, information criteria are used. They consider goodness of fit and

¹ The German version of the CTt and CTS are available from the authors upon request.

² The Rasch model predicts as the probability for a correct answer: $\exp(\text{proficiency} - \text{difficulty}) / (1 + \exp(\text{proficiency} - \text{difficulty}))$. For instance, a person with a proficiency of 1.5 Logit is expected to solve an item with a difficulty of 1.0 Logit with a probability of 62.2%.

penalize model complexity. We compare models with 1 to 6 latent classes and select the one with the lowest Akaike's information criterion (AIC), as recommended by Böhner (2011, p. 547).

Besides person homogeneity, unidimensionality is necessary to justify the allocation of students and all items on a common Logit scale. The CTt is designed to be unidimensional. Hence, we do not have any assumptions about meaningful factors, other than CT driving students' response behavior. To check for unidimensionality, we relied on confirmatory factor analysis ('lavaan 0.6–7' package in R; Rosseel, 2012). Since the data are ordinal (correct/incorrect), we applied a WLSMV-estimator (Li, 2016). A chi-square test acts as a global fit test. Furthermore, we rely on CFI, TLI, RMSEA, and SRMR as fit measures. Cut-off values for a decent fit may be: CFI and TLI > 0.95, RMSEA < 0.08, and SRMR < 0.11 (Böhner, 2011, pp. 425–427). Poor fit measures would indicate omitted factors that drive the response behavior. For example, if different answer styles were used (see Figs. 3 and 4), this could explain (besides CT) the answer behavior.

The linear logistic test model, as an extension of the Rasch model (Fischer, 1973), allows us to assess whether the cognitive operations involved in the CTt, e.g., sequencing, can explain a substantial proportion of item difficulty. This may be the prerequisite for forming proficiency levels based on the cognitive operations (AERA et al., 2014). A proportion of explained variance (R^2) of 26% might be the minimum that justifies the use of the cognitive operations for forming proficiency levels (Hartig et al., 2012). We utilize the 'eRm 1.0–1' package in R to estimate the linear logistic test model.

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; OECD, 2015, pp. 148–151). The deviance from the item discrimination implied by the Rasch model is evaluated by means of the weighted mean square error (wMNSQ = Infit). Discrimination, along these lines, means to separate students in terms of their CT proficiency. For example, if all students were able to master a certain item, this item would have no discriminatory power. The wMNSQ should lie between 0.8 and 1.2; however, wMNSQ values up to 1.33 might be acceptable (Wilson, 2005, p. 129). Items above the upper limit have a too low discrimination, whereas items below the lower limit have a too high discrimination. The point-biserial correlation is a measure for item discrimination in classical test theory and 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.

3.2.2 Building Proficiency Levels

For a criterion-referenced interpretation of the CTt results, we form proficiency levels utilizing the approach in the PISA studies (OECD, 2017, pp. 276–287). To this end, we split up the continuum of CT. We choose a width of 1.0 logits for the proficiency levels and a response probability of 62%. This means that students at the bottom of a proficiency level are expected to solve items at the bottom of that level with a probability of 62%, and at the top of the level with a probability of 38%. We opted for a width of 1.0 logits for the proficiency levels, instead of 0.8 as used in the PISA studies, because the manifestation of the cognitive operations indicates a width of one logit, which is permitted (OECD, 2017, p. 281). For every proficiency level, we provide an anchor item. These items are located about 0.5 Logit below the start of the respective level on the Wright map, corresponding with a response probability of about 62%.

3.3 RQ2: Identifying CT Profiles

3.3.1 Psychometric Test Validation

Before identifying latent profiles, we have to evaluate the psychometric properties of the CTS. Since we measured the items on a seven-point scale of rating, ranging from 'not true at all' to 'entirely true', we utilize confirmatory factor analysis with an MLR-estimator (Robitzsch, 2020). To assess the overall model fit, we use CFI, TLI, RMSEA, and SRMR. Moreover, we check for convergent and discriminant validity relying on the average variances extracted and the heterotrait–monotrait ratio. An average variance extracted greater than 0.5, and a heterotrait–monotrait ratio smaller than 0.85, indicate sufficient convergent and discriminant validity (Hair et al., 2019). An average variance extracted above 0.5 implies that more than 50% of the item variance can be explained by the corresponding factor and less than 50% is error variance. A heterotrait–monotrait ratio below 0.85 indicates that the used items capture empirically distinguishable constructs. Revell's omega total (ω) acts as a measure for internal consistency reliability because it is superior to Cronbach's alpha (α) (McNeish, 2018). Since α is widely used, however, we report it along with ω .

Analogous to the DIF-analyses for the CTt, we have to assess the measurement invariance of the CTS, i.e., a similar meaning of the constructs among subgroups, e.g., among males and females. To control for measurement invariance we apply the approach of van de Schoot et al. (2012). Since our aim is to build latent classes based on manifest means, we have to demonstrate full uniqueness measurement invariance. We use a likelihood ratio test to compare the unrestricted model where all parameters are freely estimated with a model where loadings, intercepts, and error variances across groups are restricted to be equal. In line with the DIF-analyses for the CTt, we form groups based on gender, age, computer literacy, and ability to program.

3.3.2 Identifying Latent Profiles

We use the 'tidyLPA 1.0.8' R-package in combination with MPlus 8 to identify student CT profiles by means of a latent profile analysis (LPA) (Hallquist & Wiley, 2018; Rosenberg et al., 2018). We apply an MLR-estimator (Scherer et al., 2021); missing data are not present. In light of our rather small sample size, we have to restrict variances to be equal across profiles and the covariance among the variables to be zero in order to achieve convergence (Meyer & Morin, 2016). The critical step in the LPA is to identify an appropriate number of profiles. This decision might be based on information criteria and likelihood ratio tests, as well as on conceptual deliberations (Scherer et al., 2021). Against this backdrop, we first assessed different class solutions. Following Morin and Marsh (2015) and Hofmans et al. (2020), we report the information criteria AIC, CAIC, BIC, and aBIC, as well as the bootstrap likelihood ratio test (BLRT). Since these criteria may point to a different number of optimal profiles, we also utilize the approach of Akogul and Erisoglu (2017) where information criteria are weighted to determine the optimal number of latent profiles (from an empirical point of view). The number of constructs in the LPA could be a reasonable maximum for the number of latent classes (Tondeur et al., 2019), i.e., if four constructs are considered in the LPA, four profiles could be the maximum. The identified solution should have a sufficiently high precision of classification, indicated by an entropy

greater than 0.7. However, the entropy should not be used as a model selection criterion (Sarstedt et al., 2011). To demonstrate the robustness of the findings we conduct a replication of the LPA with 100 bootstrap samples of 150 students from our overall sample of 202 students (Vanslambrouck et al., 2019). Besides this, the profiles should be substantially different from each other, which can be checked by means of a MANOVA (Tondeur et al., 2019). Afterwards, we evaluate if this approach yields a meaningful solution. The profiles should be of reasonable size and show substantial shape differences, i.e., specific strength and weakness profiles that not only differ in levels but also in their pattern (Morin & Marsh, 2015).

4 Results

4.1 RQ1: CT Proficiency Levels

4.1.1 Psychometric Validity of the CTt (German Version)

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, $p=0.11$), age ($\chi^2=16$, df=26, $p=0.94$), or computer literacy ($\chi^2=30$, df=26, $p=0.26$). However, utilizing the median of the CTt score and ability to program as a split criterion yielded significant DIF: $\chi^2=77$, df=27, $p<0.01$ and $\chi^2=48$, df=25, $p<0.01$, respectively. Four items caused this overall DIF-effect. Item 1 was far too easy for the students in our sample (-4.80 Logits) and therefore has no discriminatory power. Item 10 may have caused problems due to a different response format. The provided answer 'Options A and C are correct' might have confused students: many high-achieving students selected Option A. Items 1 and 10 can be found in Appendix 1. For items 11 and 20, we could not find a reason on the content level. Moreover, the DIF-effects are only light to moderate: Logit=0.58 and 0.51, respectively. Against this background and considering content validity in terms of alignment with the framework of Brennan and Resnick (2012), we decided to exclude items 1 and 10 from the test and retain items 11 and 20. All further analysis was carried out without items 1 and 10, i.e., with 26 items.

The mixed Rasch analysis revealed a one-class solution; the AIC is lower in comparison to any multiclass solution (e.g., AIC for one class=4944, AIC for two classes=4982, and AIC for three classes=5024).

The assumption of unidimensionality (item homogeneity) of the CTt is justified. The CFA with the 26 items loading on a single factor showed a decent fit: $\chi^2(199)=341$ ($p=0.049$), CFI=0.964, TLI=0.961, RMSEA=0.026 (90% CI [0.000, 0.039]), SRMR=0.063.

The linear logistic test model indicated the following cognitive operations as predictors of CTt item difficulty: sequencing: 1.82 Logit, 95% CI [1.46, 2.18], conditionals: 0.24 Logit, 95% CI [0.09, 0.39], functions: 0.77 Logit, 95% CI [0.54, 0.99], and variables: 2.02 Logit, 95% CI [1.77, 2.02]. Overall, these characteristics can explain 62.8% of the difficulty variance of the 26 items, which is well above the minimum acceptable value of 26%.

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, but below 1.33. All point-biserial correlations are higher than 0.30. The percentage of correct answers for all

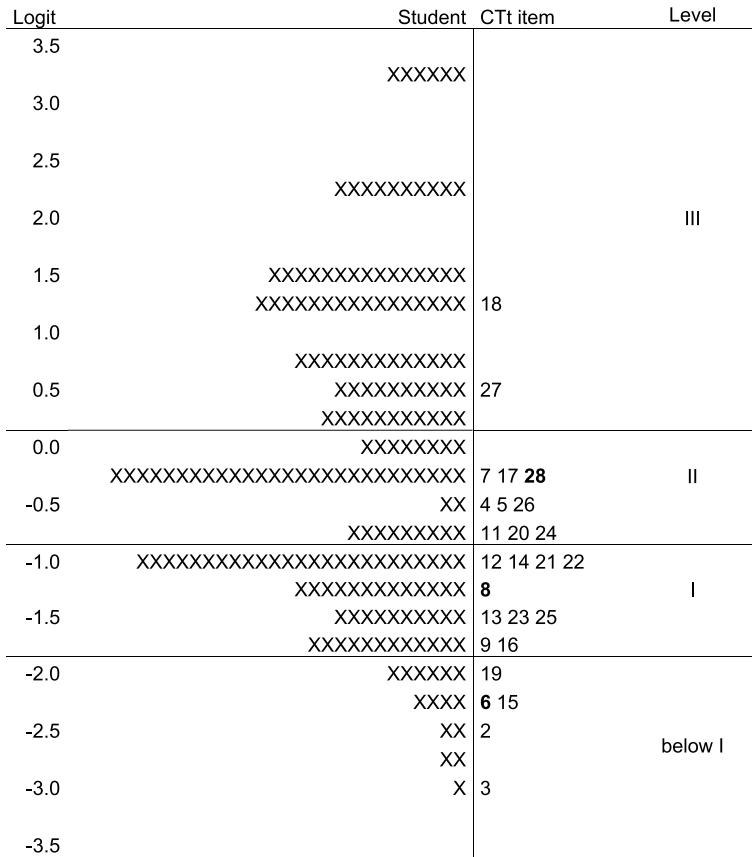


Fig. 1 Wright Map of CTt and proficiency levels

items lies between 90 and 25%. Every student fully processed the items; missing values are not present. Table 2 summarizes the item characteristics.

As the Wright map (see Fig. 1) indicates, the items are slightly too easy for the students in our sample. Nevertheless, the EAP/PV-reliability equals 0.85, WLE-reliability 0.81, which is sufficiently high for research purposes. However, if more difficult items were used, we could expect an even higher reliability.

4.1.2 Proficiency Levels

The proficiency levels are illustrated in Fig. 1. All anchor items, in bold, have negligible DIF-effects (<0.41 Logit).

Level I ($-1.75 \leq \text{Logit} < -0.75$) and below: Level I is characterized by tasks using a flow structure without further elements, like conditionals or variables. Item 6 is an anchor item for this level (see Fig. 2). Of the students in the sample, 7.4% do not achieve level I; hence, they are systematically unable to perform sequencing tasks.

For younger students, it could be necessary to insert a level below level I, which contains simple flow structures. Students in our sample, however, solved tasks like the

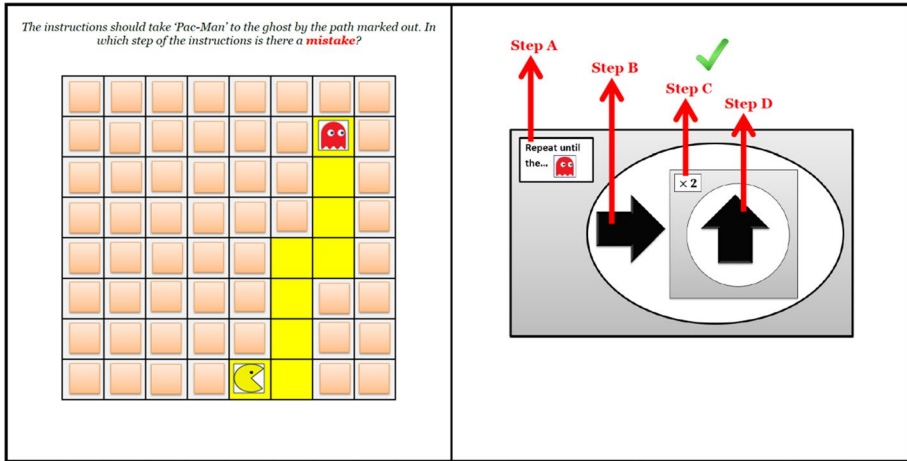


Fig. 2 Item 6, containing sequencing

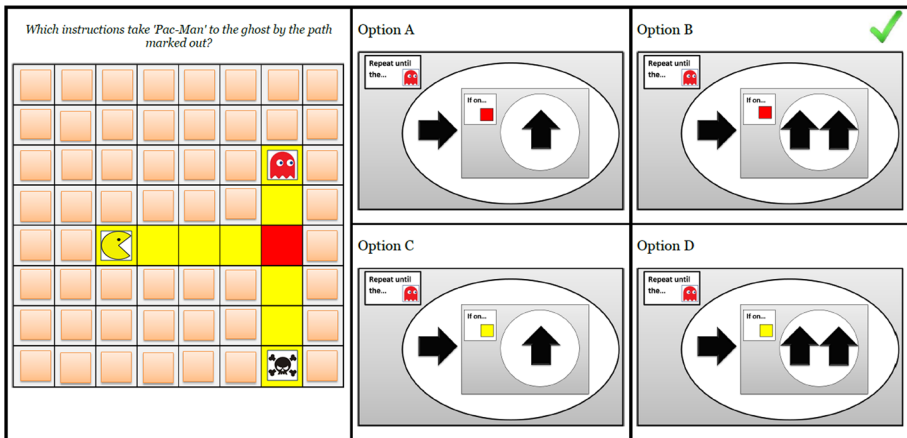


Fig. 3 Item 8, containing sequencing and conditionals

Table 1 CT dimensions (Korkmaz et al., 2017)

CT dimension	Description
Creativity	Coming up with new products; carrying out tasks in new ways, developing new ideas, finding new solutions, taking new viewpoints
Algorithmic thinking	Formalizing the solution of a problem, using a step-by-step approach to come to a solution
Cooperativity	Collaboration with others for a specific purpose
Critical thinking	Testing the reliability of information, avoiding cognitive errors, asking questions
Problem solving	Overcoming obstacles to achieve a specific goal

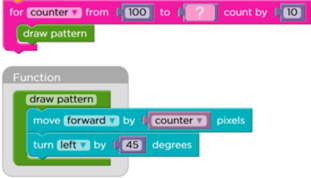

<p>The following instructions are intended to prompt the artist to draw the following pattern. 'Counter' is a variable. This variable starts at 100 and changes with each run by the step size 10 up to the searched value (?). What is the searched value?</p>  	<p>Option A 20 ✓</p> <p>Option B 50</p> <p>Option C 100</p> <p>Option D 200</p>
---	---

Fig. 4 Item 28, containing sequencing, functions, and variables.

excluded item 1 (see Appendix 1) in almost all cases. Of the students in the sample, 29.7% are on level I, i.e., they are able to systematically solve sequencing tasks.

Level II ($-0.75 \leq \text{Logit} < 0.25$): Level II comprises tasks that include conditionals and/or functions. These are core elements of CT and important in many domains. For instance, they are necessary for an effective use of spreadsheet software like Microsoft Excel. Item 8 is an anchor item for this level (see Fig. 3). In comparison to item 6, the increase in difficulty may be attributed to the use of conditionals. Of the students in the sample, 22.8% are on level II.

Level III ($0.25 \leq \text{Logit}$): On level III, items typically include the use of variables, which is also a core concept of CT (and programming). Item 28 is an anchor for level III (see Fig. 4). Of the students in the sample, 40.1% are on level III; hence, they are systematically able to cope with sequencing, conditionals, functions, and variables.

We cannot set a fourth level because we do not have meaningful item characteristics that justify building such a level, i.e., no suitable anchor items are available.

4.2 RQ2: CT Latent Profiles

4.2.1 Psychometric Validity of the CTS (German Version)

In our sample, the fit-values of the initial version of the CTS with 29 questions indicated room for improvement: $\text{YB-}\chi^2(340)=657$ ($p < 0.001$), $\text{CFI}=0.881$, $\text{TLI}=0.868$, $\text{RMSEA}=0.073$, $\text{SRMR}=0.095$. The reasons are mainly due to cross loadings. For instance, the first question 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 dimensions. This approach yielded a decent fit: $\text{YB-}\chi^2(80)=85$ ($p=0.341$), $\text{CFI}=0.997$, $\text{TLI}=0.996$, $\text{RMSEA}=0.018$ (90% CI [0.000, 0.047]), $\text{SRMR}=0.040$. Convergent and discriminant validity are fulfilled. The average variance extracted is greater than 0.543 for all five constructs. The heterotrait–monotrait ratio is smaller than 0.706 for all combination of constructs. The five dimensions are

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

# German version	Item difficulty and discrimination						DIF in Logit					
	# Original version	θ	s.e. θ	wMNSQ	Pt.bis	P+ (%)	Ability	Gender	Age	CL	Prog	
2	4	-2.42	0.23	0.99	0.34	87	-0.04	-0.09	-0.01	-0.13	0.40	
3	7	-2.75	0.25	0.99	0.30	90	-0.06	0.41	0.18	-0.30	0.40	
4	8	-0.32	0.16	1.14	0.42	55	-0.37	0.18	-0.17	0.07	-0.14	
5	10	-0.45	0.16	1.08	0.45	58	-0.20	-0.13	0.03	-0.04	-0.11	
6	11	-2.22	0.21	1.07	0.33	85	-0.07	0.17	0.09	-0.09	-0.02	
7	12	-0.18	0.16	0.97	0.54	53	0.23	0.15	0.21	-0.25	-0.10	
8	13	-1.24	0.18	1.15	0.34	72	-0.45	-0.37	-0.01	-0.06	-0.17	
9	14	-1.63	0.19	0.97	0.44	78	-0.11	0.10	0.26	-0.41	0.48	
11	16	-0.72	0.17	0.92	0.56	63	0.33	0.43	0.02	-0.06	0.58	
12	17	-0.80	0.17	1.09	0.41	64	-0.23	-0.22	-0.10	-0.07	-0.47	
13	18	-1.40	0.18	1.05	0.39	74	-0.25	-0.37	0.41	-0.25	-0.46	
14	19	-0.94	0.17	1.07	0.42	67	-0.23	-0.03	0.03	0.17	-0.23	
15	20	-2.01	0.20	0.91	0.45	83	0.38	0.10	-0.04	0.32	-0.01	
16	21	-1.53	0.18	0.98	0.46	76	0.05	0.35	-0.33	0.21	0.17	
17	22	-0.18	0.16	0.94	0.57	53	0.05	0.26	-0.16	0.18	-0.26	
18	23	1.45	0.19	1.22	0.30	25	-0.58	-0.13	0.33	0.08	-0.27	
19	24	-1.85	0.20	0.92	0.47	81	0.16	0.03	-0.03	0.16	0.21	
20	25	-0.69	0.17	0.89	0.58	62	0.51	0.32	-0.06	0.02	0.03	
21	26	-0.94	0.17	0.94	0.53	67	0.38	-0.25	-0.11	0.26	-0.30	
22	27	-0.91	0.17	1.04	0.46	66	0.04	-0.17	0.04	0.00	-0.48	
23	28	-1.40	0.18	1.01	0.45	74	-0.18	-0.31	-0.18	0.02	-0.06	
24	-	-0.50	0.16	0.89	0.59	59	0.14	-0.12	-0.19	0.33	0.14	
25	-	-1.46	0.18	0.91	0.52	75	0.29	-0.29	-0.18	0.08	0.21	
26	-	-0.26	0.16	0.93	0.57	54	0.20	-0.03	0.12	-0.17	0.05	
27	-	0.51	0.17	0.96	0.56	40	0.00	0.05	-0.12	0.09	0.02	

Table 2 (continued)

# German version	Item difficulty and discrimination				DIF in Logit						
	# Original version	θ	s.e. θ	wMNSQ	Pt.bis	P+ (%)	Ability	Gender	Age	CL	Prog
28	-	-0.10	0.16	0.97	0.55	51	0.03	-0.01	-0.02	-0.16	0.41

θ item difficulty in Logit, s.e. standard error of item difficulty, wMNSQ weighted mean square error, Pt.bis. point biserial correlation, P+ percentage of correct responses, Ability median of CIt raw score, CL computer literacy (INCOBI-R), Prog. programming ability (yes/no)

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

Construct	Item	Mean (SD)	λ	α	ω	AVE	Latent correlations below diagonal, square root of AVE on diagonal, manifest correlations above diagonal				
							(1)	(2)	(3)	(4)	(5)
(1) Creativity	cr_3	5.6 (1.5)	0.75	0.87	0.87	0.75	0.87	0.28	0.19	0.59	0.17
	cr_4	5.6 (1.3)	0.86								
	cr_5	5.3 (1.3)	0.89								
(2) Algorithmic thinking	al_3	4.0 (1.9)	0.90	0.90	0.90	0.73	0.32	0.85	0.13	0.53	-0.00
	al_4	3.7 (1.8)	0.86								
	al_5	4.0 (1.8)	0.83								
(3) Cooperativity	co_1	4.6 (1.8)	0.85	0.89	0.89	0.70	0.22	0.15	0.84	0.19	-0.14
	co_2	4.2 (1.8)	0.90								
	co_3	4.8 (1.7)	0.82								
(4) Critical thinking	cr_1	5.1 (1.4)	0.74	0.80	0.80	0.57	0.71	0.62	0.22	0.75	0.13
	cr_2	4.9 (1.4)	0.75								
	cr_3	4.6 (1.4)	0.78								
(5) Problem solving	pr_1	5.5 (1.5)	0.79	0.78	0.78	0.54	0.20	-0.03	-0.20	0.15	0.73
	pr_2	5.4 (1.7)	0.68								
	pr_4	5.1 (1.5)	0.74								

Items measured on a 7-point rating scale. λ =standardized factor loading, α Cronbach's alpha, ω Revell's omega total, AVE average variance extracted. Figures in bold indicate significant correlations at the 5% level

Table 4 Information criteria, entropies and BLRT results for one to five latent profiles

No. profiles	-LL	No. Par	AIC	CAIC	BIC	aBIC	Entropy	p(BLRT)
1	1421	8	2859	2893	2885	2859	1.000	–
2	1360	13	2747	2803	2790	2749	0.734	0.000
3	1334	18	2704	2782	2764	2707	0.744	0.000
4	1321	23	2687	2786	2763	2690	0.763	0.000
5	1314	28	2680	2800	2772	2684	0.786	0.286

-LL -Log-likelihood, *No. Par.* number of estimated parameters, *AIC* Akaike's information criterion, *CAIC* Consistent AIC, *BIC* Bayesian Information Criterion, *aBIC* adjusted BIC, *BLRT* bootstrap likelihood ratio test

reliably measured (α and $\omega > 0.77$). The characteristics of the refined version of the CTS can be found in Table 3; the questions can be seen in Appendix 2.

Full uniqueness measurement invariance is ensured for the four considered subgroups: gender: $\Delta\chi^2(105) = 121.84$, $p = 0.125$; age: $\Delta\chi^2(105) = 116.98$, $p = 0.200$; computer literacy: $\Delta\chi^2(105) = 122.92$, $p = 0.112$; and ability to program: $\Delta\chi^2(105) = 128.94$, $p = 0.056$. Hence, it may be justified to use manifest means for the LPA.

4.2.2 Latent Profiles

The descriptive statistics in Table 3 show that, on average, the students assess their creative thinking, algorithmic thinking, cooperativity, critical thinking, and problem solving above the neutral scale mean ($=4$). The highest latent correlation appears between creativity and critical thinking ($\rho = 0.71$, $p < 0.001$). However, there are also small and statistically insignificant correlations, e.g., between algorithmic thinking and cooperativity ($\rho = 0.15$, $p = 0.079$).

When identifying latent profiles, the dimension *problem solving* was problematic. The likely reason is the reverse coding of the corresponding items; see Appendix 2. Including this dimension in the LPA yielded spurious profiles, e.g., a profile with students who score very low in problem solving and very high in all other dimensions. Since considering this construct could bias the findings, we removed it from the further analysis. This may also be suitable from a conceptual point of view. Creativity, algorithmic thinking, critical thinking, and cooperativity might all be necessary during the course of problem solving; problem solving might be a construct of a different nature.

Table 4 depicts the information criteria, the BLRT results, and the entropies. Based on the information criteria AIC and aBIC, a five-profile solution would be optimal; CAIC, BIC, and BLRT, as well as the analytic hierarchy process of Akogul and Erisoglu (2017), point to four profiles. Moreover, the entropy for this solution is sufficiently high (0.763). A replication of the LPA with 100 bootstrap samples of 150 students lends support to a four-class solution. In 2% of the cases, three profiles are optimal; in 30% of the cases, four profiles; and in 68% of the cases, five profiles. An inspection of the five-class solutions, however, revealed problematic profiles, e.g., profiles with only two students. Since four profiles may be the maximum from a conceptual point of view and three or less profiles seem to be insufficient, four latent profiles could be the appropriate number.

The MANOVA yielded significant different means among the four profiles: $F(12, 516) = 69.90$, Wilk's $\Lambda = 0.078$, $p < 0.001$, $\eta^2 = 0.573$. In other words, 57.3% of the variance

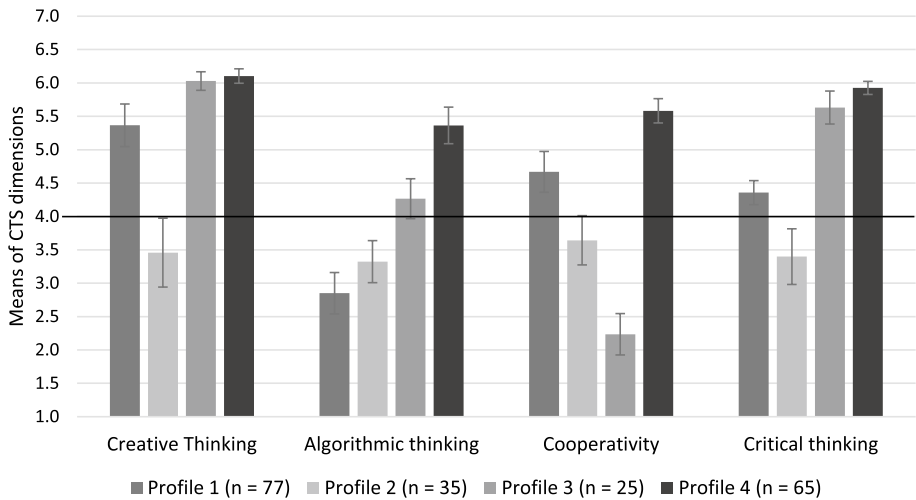


Fig. 5 Description of the four latent profiles based on CTS (N = 202)

can be explained by profile membership. These findings and Fig. 5 point to sufficiently distinct profiles.

The identified profiles may be meaningful from a conceptual point of view:

- Profile 1 (creative thinkers with a focus on collaboration): This profile comprises 77 (38.1%) students and is the largest one. Students in this profile perceive their creative thinking to be well above the neutral scale mean. Moreover, they also assess their cooperativity and critical thinking, on average, as above the neutral scale mean. However, their perceived algorithmic thinking skills are rather low.
- Profile 2 (low-level computational thinkers): These 35 (17.3%) students, on average, self-assess all dimensions of CT as below the scale mean, i.e., they consider themselves as not being capable of performing sufficiently well in the four CT dimensions.
- Profile 3 (computational thinkers with low cooperativity): This profile consists of 25 (12.4%) students, which is the smallest profile. Overall, the students self-assess their creative and critical thinking as very high. Their algorithmic thinking is, on average, slightly above the scale mean. However, these students report very low cooperativity.
- Profile 4 (high-level computational thinkers): Sixty-five (32.2%) students belong to this profile and report high levels across all four CT dimensions.

4.3 RQ3: Relationship Between CTt and CTS Results

The latent correlations between the CTt and the five dimensions of the CTS are with *creativity* 0.271 ($p=0.002$), *algorithmic thinking* 0.309 ($p<0.001$), *cooperativity* -0.003 ($p=0.956$), *critical thinking* 0.408 ($p<0.001$), and *problem solving* 0.154 ($p=0.085$). Considering all CTS dimensions as independent variables, and the CTt as a dependent variable in a latent regression, only *algorithmic thinking* is statistically significant ($b=0.319$ Logit, $p<0.001$). Figure 6 presents the relationship between the three proficiency levels (CTt) and the four latent profiles (CTS). The notches in the boxplots represent the 95% confidence intervals. As can be seen, students in profiles 1 and 2 score significantly lower

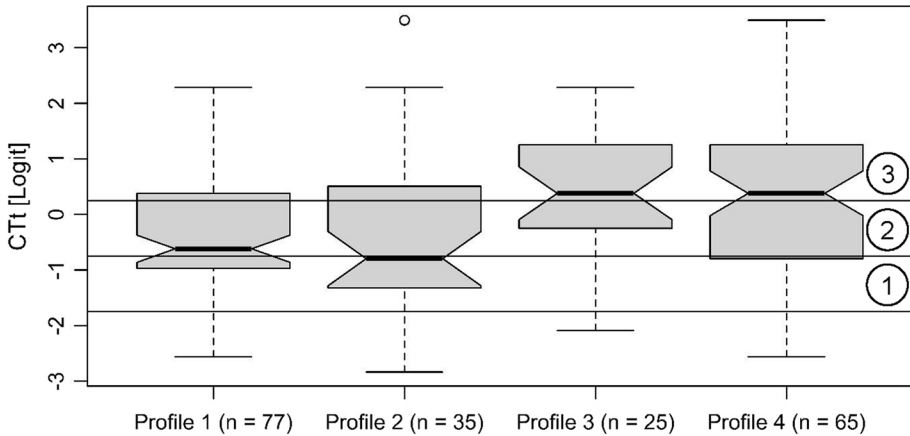


Fig. 6 CTt proficiency in Logit and proficiency levels (1, 2, 3) grouped by latent profile membership (CTS). Notches indicate 95% confidence intervals

in the CTt than students in profiles 3 and 4. The median proficiency in profiles 1 and 2 is around the threshold of proficiency level II, and in profiles 3 and 4, around the threshold of proficiency level III.

5 Discussion, Limitations, and Outlook

5.1 Discussion

The aim of this paper was to contribute to more vivid interpretation of CT assessment findings. To this end, we relied on a performance test (CTt) and a self-assessment instrument (CTS). Both are freely available and internationally accepted standardized instruments (Shute et al., 2017).

In terms of the CTt (RQ1), the main advantage is its reliance on the well-established framework of Brennan and Resnick (2012). Furthermore, the CTt details the cognitive operations that underlie the test items. This may be the basis for a theoretically founded interpretation of the test results. Utilizing IRT, we were able to demonstrate Rasch scalability, which implies the specific objectivity of the test. This is the basis for a criterion-referenced test interpretation. The presented proficiency level model for CT is an additional step in capturing the construct of CT: it adds a model for describing item difficulty and cognitive demand. A proficiency level model could help to communicate the test results. A statement like ‘Student A is able to solve tasks that contain sequencing and conditionals’ is more informative than ‘Student A solved 16 out of 26 items correctly’. It could also help teachers to better tailor their instruction to specific student needs. For instance, a student on proficiency level II might not be able to correctly solve items with variables, but is able to do so as regards items with conditionals and functions. This insight can be utilized for scaffolding processes. Before designing instructional measures, it may be important to evaluate students’ prior knowledge (Bransford et al., 2000). The CTt can carry out this purpose as it does not require specific prior knowledge, such as a programming language.

Concerning the CTS (RQ2), our confirmatory factor analysis indicated, in line with Korkmaz et al. (2017), that five dimensions can be identified. Due to substantial cross-loadings, we had to exclude items from the original version, which contained 29 items. However, the author of this initial version also reduced the initial scale to 20 items in a later study (Korkmaz & Bai, 2019). By means of an LPA, we identified four profiles.

Concerning specific strengths and weaknesses, profiles 1 and 3 are of special interest. Students in profile 1 report a weakness in algorithmic thinking. Algorithmic thinking can be regarded as the link between CT and computer science (Doleck et al., 2017) and may be a technical skill. To foster students' algorithmic thinking skills, it may be preferable to rely on visual programming languages because professional programming languages can be regarded as difficult or boring (Lye & Koh, 2014; Repenning, 2017). Scratch could be a suitable option for this purpose. Grover et al. (2015) demonstrated the effectiveness of Scratch in fostering algorithmic thinking. Moreover, scalable game design (Repenning, 2018) could be a viable option to improve students' algorithmic thinking. The advantage is that students start with a project instead of first being confronted with code and syntax. Students can bring in their creative and critical thinking skills and afterwards learn technical aspects, which keeps them in the zone of proximal flow (Repenning et al., 2015). Besides this, teachers could form groups where students in profile 1 collaboratively learn with students from other profiles, especially students in profile 4 who have high levels of perceived algorithmic thinking skills. Students in profile 1 can bring in their creativity and critical thinking and benefit from the other students' high algorithmic thinking skills.

Profile 3 is remarkable because these students report low cooperativity and, at the same time, high creative and critical thinking. Colloquially, these students might be referred to as nerds. For these students it could be important to increase their cooperativity, which is regarded as a core 21st-century skill (van Laar et al., 2017). Again, Scratch could be a promising approach because it relies on a social learning paradigm and allows members of the community to learn from each other through the opportunity to share and extend projects (Jiang et al., 2021; Repenning et al., 2015; Resnick & Rusk, 2020; Shute et al., 2017). Students in profile 3 could become aware of the benefits of collaboration when working on Scratch projects. Teachers may put a special focus on students in this profile in order to integrate them into the group and facilitate teamwork.

Overall, 67.8% of the students in our sample might require specific support. Students in CT profile 1 may need help with increasing their algorithmic thinking skills. Students in profile 2 score low in all four CTS dimensions. Students in profile 3 seem to have a deficit in cooperativity. Only students in profile 4 report high values among all four CTS dimensions. Teachers could make use of these students to support fellow students with deficits in one or more CT dimension. In general, collaboration seems to be conducive for fostering CT (Denner et al., 2014).

Concerning the relationship between the CTt and CTS results (RQ3), the correlations found in our study are well in line with the correlations reported by Polat et al. (2021), indicating that the findings might be robust across various populations. Moreover, we show that when considering all CTS dimensions as independent variables and the CTt as the dependent variable in a latent regression, only algorithmic thinking is a statistically significant predictor. Hence, the CTt may primarily be related with algorithmic thinking of the ISTE framework. This can also be seen from Figs. 5 and 6. Students in profiles 3 and 4, with the highest reported algorithmic thinking, achieve significantly higher CTt scores than students in profiles 1 and 2, and reach higher proficiency levels.

Since the CTt and CTS rely on different methods, i.e., performance test and self-assessment, it is likely that constructs of a different nature are captured. Self-assessments

might capture self-efficacy beliefs (Scherer et al., 2017), which play an important role in predicting (intended) behavior (Fishbein & Ajzen, 2010). For instance, students' perceived cooperativity may be a good predictor for their actual collaboration in CT projects. Moreover, self-assessments are very cost-efficient (Scherer et al., 2017). The CTS requires less than five minutes of test time. Concerning the identification of latent profiles, self-assessment instruments seem to be an established method (e.g., Scherer et al., 2021). Overall, self-assessments may not be inferior to performance tests.

If the purpose is to investigate the nomological net of CT, however, performance tests may be more suitable than self-assessments. Polat et al. (2021) investigated the relationship of CT with the external variables of mathematics and information technologies course achievement. The correlations are substantially higher when the CTt instead of the CTS is used. Moreover, forming proficiency levels based on self-assessments may not be a viable option.

Overall, the CTt and CTS may well complement each other. If we had only used the CTt, the remarkable CT profiles 1 and 3 would not have been detected; students in these profiles may need specific attention.

5.2 Limitations

Our study is not without limitations. A general limitation is that our sample is narrow in scope as it comprises only students from German-speaking Switzerland and from one type of school (high school). Concerning RQ1 and the use of the CTt, one disadvantage is its reliance on dichotomous constructed response items. It may not be able to capture higher-level computational concepts. The formed proficiency levels can be interpreted in a meaningful way because they are linked to previously specified characteristics, e.g., sequencing. However, for high-achieving students, items that cover more advanced computational concepts would be necessary. We will come back to this point in the outlook section below.

Concerning RQ2 and the use of the CTS, we had to exclude items from the original version due to a lack of discriminant validity. On the one hand, this exclusion contributes to the psychometric validity of the instrument and content validity may still be achieved; on the other hand, however, results based on the full version of the CTS could be difficult to compare with our refined version. Besides this, we did not consider the CTS dimension *problem solving* in the LPA. The reason for this was that all the items that operationalize problem solving are reversely coded; consideration of the dimension in the LPA yielded spurious profiles. Moreover, due to our rather small sample size, we had to use a parsimonious LPA model where equal variances across the profiles and covariances of zero are assumed. These assumptions are restrictive (Scherer et al., 2021). Moreover, we used manifest means as the basis for the LPA. This may be unavoidable due to our sample size but it neglects measurement error (Meyer & Morin, 2016). As a robustness check, we also used regression factor scores instead of manifest means; it yielded similar results (see Appendix 3). Against the backdrop of our cross-sectional sample of high school students, it is doubtful whether the identified four profiles can be replicated in samples with younger students and be consistent across time (Meyer & Morin, 2016). Moreover, the identification of the profiles was exploratory, which may be inherent to the LPA but not ideal from a theoretical point of view (Hofmans et al., 2020). Based on our simulation approach, however, we may conclude that four profiles could be a reasonable minimum.

5.3 Outlook

The usefulness of the CTt for assessing students with a CT proficiency equal or higher to those in our sample could be increased. To this end, it would be necessary to set further meaningful proficiency levels. In order to anchor these proficiency levels, items of greater difficulty should be constructed on a criterion basis. Such a criterion could be the CT concept *diffusion* (Repenning, 2017, pp. 18–19). Diffusion goes beyond conditionals, functions, and variables. It is a kind of artificial intelligence and can, for instance, be used to move objects dynamically. As can be seen from the tasks depicted in Figs. 2 and 3, instructions are required to lead Pacman to the ghost. However, if Pacman should autonomously find the ghost, utilizing diffusion is a viable option (see Repenning, 2006). It can be claimed that the ghost has a ‘scent’ that spreads in the maze. This process is modeled by means of diffusion equations. Pacman can then use a hill-climbing approach to find the ghost. It checks the scent concentration in all four neighboring fields and moves to the field with the highest concentration. This is an efficient approach because it takes obstacles like walls into account. To test students’ understanding of this concept, they could be asked how long it will take, depending on the speed of diffusion, to detect the ghost. They may also predict the path Pacman will choose to reach the ghost or implement the hill-climbing approach using visual code blocks. We think including diffusion into the CTt is promising because it is an important concept in many domains (Repenning, 2006). Examples are the diffusion of heat (physics) or osmosis (biology).

In future research, it could be promising to use performance tests to capture the five dimensions covered by the CTS and form proficiency levels. Suitable tests are available and could be adapted for CT: for creativity (Israel-Fishelson & Hershkovitz, 2022, Appendix C), for algorithmic thinking (Román-González et al., 2017), for cooperativity (Salas et al., 2017), for (complex) problem solving (Greiff et al., 2013), for collaborative problem solving (Stadler et al., 2020), and for critical thinking (Ennis, 1993). If a self-assessment instrument should be used, the computing attitudes survey (Dorn & Tew, 2015) could be a viable alternative to the CTS.

6 Conclusion

This paper contributes to a better understanding of computational thinking as a construct. Based on a proficiency level model, distinct levels of computational thinking performance can be described, i.e., characteristics of computational thinking tasks that students on a specific level are systematically able to master but which cannot be mastered by students on a lower level. We formed proficiency levels based on the Computational Thinking Test, which covers the cognitive operations (computational concepts) of sequencing, conditionals, functions, and variables. Moreover, we identified latent profiles based on the Computational Thinking Scale. Our findings indicated that 67.8% of the students may need specific guidance and support. Students in the first profile (38.1%) reported deficits in algorithmic thinking while students in the second profile (17.3%) self-assessed themselves as low in all four dimensions: creative thinking, algorithmic thinking, cooperativity, and critical thinking. The third profile (12.4%) comprises students who were reluctant to cooperate. Knowledge about the identified proficiency levels and the four distinctive computational thinking profiles could help teachers offer person-centered guidance and support to their students.

Appendix 1: Excluded items due to psychometric validations

Item 1. Excluded because too easy for the test takers:

Which instructions take 'Pac-Man' to the ghost by the path marked out?

<p>Option A</p>	<p>Option B</p>
<p>Option C</p>	<p>Option D</p>

Item 10. Excluded because test takers were confused by option D:

What is missing in the instructions below to take 'Pac-Man' to the ghost by the path marked out?

Repeat until the... (ghost icon) → if on... (blue square) → if on... (red square) → ↑

Option A	
Option B	
Option C	
Option D	Both option A and option C are correct

Appendix 2: Refined version of CTS

Creativity

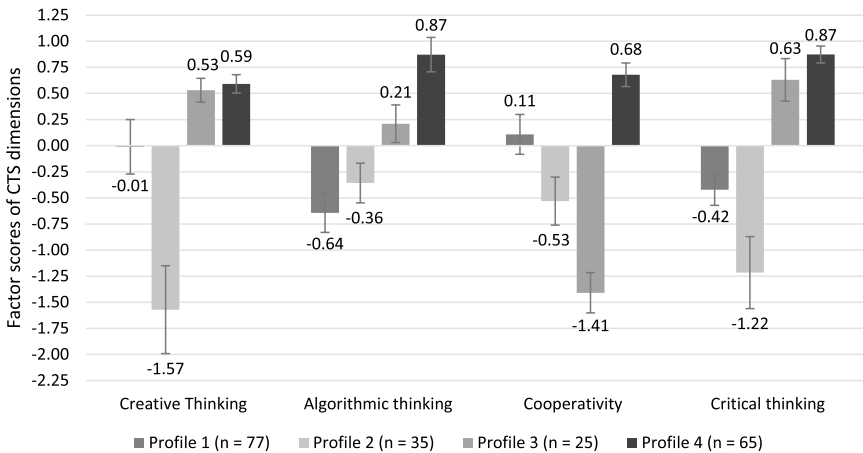
cr_3 "I believe that I can solve most of the problems I face if I have a sufficient amount of time and if I show effort."

cr_4 "I believe that I can solve possible problems that may occur when I encounter a new situation."

	cr_5	"I trust that I can apply a plan, at the same time as making it, in order to solve a problem."
Algorithmic thinking	al_3	"I think that I am better able to learn instructions with the help of mathematical symbols and concepts."
	al_4	"I can mathematically express the solutions for the problems I face in daily life."
	al_5	"I can digitize a mathematical problem expressed verbally."
Cooperativity	co_1	"I like experiencing cooperative learning together in my group of friends."
	co_2	"In cooperative learning, I think that I attain/will attain more successful results because I am working in a group."
	co_3	"I like solving problems related to a group project together with my friends in cooperative learning."
Critical thinking	cr_1	"I am willing to learn challenging things."
	cr_2	"I am proud of being able to think with great precision."
	cr_3	"I make use of a systematic method while comparing the options at hand and while reaching a decision."
Problem solving	pr_1	"I have problems in demonstrating the solution to a problem in my mind." (R)
	pr_2	"I have difficulties regarding the issue of where and how I should use variables such as X and Y in the solution of a problem." (R)
	pr_4	"I cannot apply the solutions I plan respectively and gradually." (R)

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

Appendix 3



Note. Latent profiles based on factor scores (M = 0, SD = 1) are provided to facilitate comparison across studies.

Funding The first author received a postdoctoral fellowship for carrying out this research from the Basic Research Fund at the University of St.Gallen (no. 1031542).

Data availability Readers who wish to examine the data may contact the corresponding author.

Declarations

Conflict of interest No known conflicts of interest associated with this manuscript.

Ethical standards The procedures performed in this study follow the standards of the Ethics Committee of the University of St.Gallen.

References

- AERA, APA, & NCME. (2014). Standards for educational and psychological testing. American Educational Research Association. <https://eric.ed.gov/?id=ED565876>
- Akogul, S., & Erisoglu, M. (2017). An approach for determining the number of clusters in a model-based cluster analysis. *Entropy*, *19*(9), 452. <https://doi.org/10.3390/e19090452>
- Andersen, E. B. (1973). A goodness of fit test for the Rasch model. *Psychometrika*, *38*(1), 123–140. <https://doi.org/10.1007/BF02291180>
- Barr, D., Harrison, J., & Conery, L. (2011). Computational thinking: A digital age skill for everyone. *Learning & Leading with Technology*, *38*(6), 20–23. <https://eric.ed.gov/?id=EJ918910>
- 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). ACM. <https://doi.org/10.1145/3137065.3137069>
- Bransford, J., Brown, A. L., & Cocking, R. R. (2000). *How people learn: Brain, mind, experience, and school* (Expanded edition). National Academy Press. <https://doi.org/10.17226/9853>
- 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. <https://www.media.mit.edu/publications/new-frameworks-for-studying-and-assessing-the-development-of-computational-thinking/>
- Bühner, M. (2011). Einführung in die Test- und Fragebogenkonstruktion [Introduction to test and questionnaire construction] (3rd ed.). Pearson Studium. <https://d-nb.info/1007197234>
- 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. <https://doi.org/10.3102/0034654317710096>
- Carretero, S., Vuorikari, R., & Punie, Y. (2017). DigComp 2.1: The Digital Competence Framework for Citizens with eight proficiency levels and examples of use. European Union (EU). <https://doi.org/10.2760/38842>
- Chan, S.-W., Looi, C.-K., & Sumintono, B. (2020). Assessing computational thinking abilities among Singapore secondary students: A Rasch model measurement analysis. *Journal of Computers in Education*, *8*, 213–236. <https://doi.org/10.1007/s40692-020-00177-2>
- Chen, G., Shen, J., Barth-Cohen, L., Jiang, S., Huang, X., & Eltouky, M. (2017). Assessing elementary students' computational thinking in everyday reasoning and robotics programming. *Computers & Education*, *109*, 162–175. <https://doi.org/10.1016/j.compedu.2017.03.001>
- Denner, J., Werner, L., Campe, S., & Ortiz, E. (2014). Pair programming: Under what conditions is it advantageous for middle school students? *Journal of Research on Technology in Education*, *46*(3), 277–296. <https://doi.org/10.1080/15391523.2014.888272>
- Doleck, T., Bazalais, P., Lemay, D. J., Saxena, A., & Basnet, R. B. (2017). Algorithmic thinking, cooperativity, creativity, critical thinking, and problem solving: Exploring the relationship between computational thinking skills and academic performance. *Journal of Computers in Education*, *4*(4), 355–369. <https://doi.org/10.1007/s40692-017-0090-9>
- Dorn, B., & Tew, A. E. (2015). Empirical validation and application of the computing attitudes survey. *Computer Science Education*, *25*(1), 1–36. <https://doi.org/10.1080/08993408.2015.1014142>
- Durak, H. Y., & Saritepeci, M. (2018). Analysis of the relation between computational thinking skills and various variables with the structural equation model. *Computers & Education*, *116*, 191–202. <https://doi.org/10.1016/j.compedu.2017.09.004>

- Embretson, S. E., & Daniel, R. C. (2008). Understanding and quantifying cognitive complexity level in mathematical problem solving items. *Psychology Science Quarterly*, 50(3), 328–344. <https://www.semanticscholar.org/paper/Understanding-and-Quantifying-Cognitive-Complexity-Embretson-Daniel/29a75e2f74f0d5a870a846660c617a2273b73975>
- Ennis, R. H. (1993). Critical thinking assessment. *Theory into Practice*, 32(3), 179–186. <https://doi.org/10.1080/00405849309543594>
- Ericson, B., & McKlin, T. (2012). Effective and sustainable computing summer camps. In L. Smith King, D. R. Musicant, T. Camp, & P. Tymann (Eds.), *Proceedings of the 43rd ACM technical symposium on Computer Science Education - SIGCSE '12* (p. 289). ACM Press. <https://doi.org/10.1145/2157136.2157223>
- Fischer, G. H. (1973). The linear logistic test model as an instrument in educational research. *Acta Psychologica*, 37(6), 359–374. [https://doi.org/10.1016/0001-6918\(73\)90003-6](https://doi.org/10.1016/0001-6918(73)90003-6)
- Fishbein, M., & Ajzen, I. (2010). Predicting and changing behavior: The reasoned action approach. Taylor and Francis. <https://psycnet.apa.org/record/2009-17267-000>
- Fraillon, J., Ainley, J., Schulz, W., Friedman, T., & Duckworth, D. (2019). Preparing for life in a digital world: IEA International Computer and Information Literacy Study 2018 international report. International Association for the Evaluation of Educational Achievement (IEA). <https://www.iea.nl/publications/study-reports/preparing-life-digital-world>
- Fraillon, J., Ainley, J., Schulz, W., Friedman, T., & Duckworth, D. (2020). IEA International Computer and Information Literacy Study 2018: Technical Report. International Association for the Evaluation of Educational Achievement (IEA). <https://www.iea.nl/publications/technical-reports/icils-2018-technical-report>
- Greiff, S., Wüstenberg, S., Molnár, G., Fischer, A., Funke, J., & Csapó, B. (2013). Complex problem solving in educational contexts—Something beyond g: Concept, assessment, measurement invariance, and construct validity. *Journal of Educational Psychology*, 105(2), 364–379. <https://doi.org/10.1037/a0031856>
- Grover, S., & Pea, R. (2013). Computational thinking in K-12: A review of the state of the field. *Educational Researcher*, 42(1), 38–43. <https://doi.org/10.3102/0013189X12463051>
- Grover, S., Pea, R., & Cooper, S. (2015). Designing for deeper learning in a blended computer science course for middle school students. *Computer Science Education*, 25(2), 199–237. <https://doi.org/10.1080/08993408.2015.1033142>
- Grover, S., Pea, R., & Cooper, S. (2016). Factors influencing computer science learning in middle school. In C. Alphonse, J. Tims, M. Caspersen, & S. Edwards (Eds.), *Proceedings of the 47th ACM Technical Symposium on Computing Science Education* (pp. 552–557). ACM. <https://doi.org/10.1145/2839509.2844564>
- Guggemos, J. (2021). On the predictors of computational thinking and its growth at the high-school level. *Computers & Education*. <https://doi.org/10.1016/j.compedu.2020.104060>
- Guggemos, J., Seufert, S., & Román-González, M. (2019). Measuring computational thinking - Adapting a performance test and a self-assessment instrument for German-speaking countries. *Proceedings of the 16th International Conference Cognition and Exploratory Learning in the Digital Age (CELDA)* (pp. 183–191). https://doi.org/10.33965/celda2019_201911L023
- Hair, J. F., Risher, J. J., Sarstedt, M., & Ringle, C. M. (2019). When to use and how to report the results of PLS-SEM. *European Business Review*, 31(1), 2–24. <https://doi.org/10.1108/EBR-11-2018-0203>
- Hallquist, M. N., & Wiley, J. F. (2018). MplusAutomation: An R package for facilitating large-scale latent variable analyses in Mplus. *Structural Equation Modeling: A Multidisciplinary Journal*, 25(4), 621–638. <https://doi.org/10.1080/10705511.2017.1402334>
- Harteis, C., Goller, M., & Caruso, C. (2020). Conceptual change in the face of digitalization: Challenges for workplaces and workplace learning. *Frontiers in Education*. <https://doi.org/10.3389/educ.2020.00001>
- Hartig, J., & Frey, A. (2013). Sind modelle der item-response-theorie (IRT) das „Mittel der Wahl“ für die Modellierung von Kompetenzen? [Benefits and limitations of modeling competencies by means of Item Response Theory (IRT)]. *Zeitschrift Für Erziehungswissenschaft*, 16(S1), 47–51. <https://doi.org/10.1007/s11618-013-0386-0>
- Hartig, J., Frey, A., Nold, G., & Klieme, E. (2012). An application of explanatory item response modeling for model-based proficiency scaling. *Educational and Psychological Measurement*, 72(4), 665–686. <https://doi.org/10.1177/0013164411430707>
- Hofmans, J., Wille, B., & Schreurs, B. (2020). Person-centered methods in vocational research. *Journal of Vocational Behavior*, 118, 103398. <https://doi.org/10.1016/j.jvb.2020.103398>

- Hooshyar, D., Malva, L., Yang, Y., Pedaste, M., Wang, M., & Lim, H. (2021). An adaptive educational computer game: Effects on students' knowledge and learning attitude in computational thinking. *Computers in Human Behavior*, *114*, 106575. <https://doi.org/10.1016/j.chb.2020.106575>
- 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. <https://doi.org/10.1016/j.compedu.2018.07.004>
- Ifenthaler, D., Hofhues, S., Egloffstein, M., & Helbig, C. (Eds.). (2021). *Digital Transformation of Learning Organizations*. Springer. <https://doi.org/10.1007/978-3-030-55878-9>
- Ilic, U., Haseski, H. I., & Tugtekin, U. (2018). Publication trends over 10 years of computational thinking research. *Contemporary Educational Technology*, *9*(2), 131–153. <https://doi.org/10.30935/cet.414798>
- Israel, M., Pearson, J. N., Tapia, T., Wherfel, Q. M., & Reese, G. (2015). Supporting all learners in school-wide computational thinking: A cross-case qualitative analysis. *Computers & Education*, *82*, 263–279. <https://doi.org/10.1016/j.compedu.2014.11.022>
- Israel-Fishelson, R., & Hershkovitz, A. (2022). Studying interrelations of computational thinking and creativity: A scoping review (2011–2020). *Computers & Education*, *176*, 104353. <https://doi.org/10.1016/j.compedu.2021.104353>
- ISTE. (2015). Computational thinking: leadership toolkit. <https://www.iste.org/computational-thinking>
- Jiang, B., Zhao, W., Gu, X., & Yin, C. (2021). Understanding the relationship between computational thinking and computational participation: A case study from Scratch online community. *Educational Technology Research and Development*, *69*(5), 2399–2421. <https://doi.org/10.1007/s11423-021-10021-8>
- Kirschner, P. A., & Stoyanov, S. (2020). Educating youth for nonexisting/not yet existing professions. *Educational Policy*, *34*(3), 477–517. <https://doi.org/10.1177/0895904818802086>
- Köhler, C., & Hartig, J. (2017). Practical significance of item misfit in educational assessments. *Applied Psychological Measurement*, *41*(5), 388–400. <https://doi.org/10.1177/0146621617692978>
- Kong, S. C., & Wang, Y. Q. (2021). Item response analysis of computational thinking practices: Test characteristics and students' learning abilities in visual programming contexts. *Computers in Human Behavior*, *122*, 106836. <https://doi.org/10.1016/j.chb.2021.106836>
- 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. <https://doi.org/10.1016/j.chb.2017.01.005>
- Korkmaz, Ö., & Bai, X. (2019). Adapting computational thinking scale (CTS) for Chinese high school students and their thinking scale skills level. *Participatory Educational Research*, *6*(1), 10–26. <https://doi.org/10.17275/per.19.2.6.1>
- Li, C.-H. (2016). Confirmatory factor analysis with ordinal data: Comparing robust maximum likelihood and diagonally weighted least squares. *Behavior Research Methods*, *48*(3), 936–949. <https://doi.org/10.3758/s13428-015-0619-7>
- Kong, S. C. (2019). Components and methods of evaluating computational thinking for fostering creative problem-solvers in senior primary school education. In S. C. Kong & H. Abelson (Eds.), *Computational thinking education* (pp. 119–141). Springer. https://doi.org/10.1007/978-981-13-6528-7_8
- Leys, C., Delacre, M., Mora, Y. L., Lakens, D., & Ley, C. (2019). How to classify, detect, and manage univariate and multivariate outliers, with emphasis on pre-registration. *International Review of Social Psychology*. <https://doi.org/10.5334/irsp.289>
- Lohr, A., Stadler, M., Schultz-Pernice, F., Chernikova, O., Sailer, M., Fischer, F., & Sailer, M. (2021). On powerpointers, clickerers, and digital pros: Investigating the initiation of digital learning activities by teachers in higher education. *Computers in Human Behavior*, *119*, 106715. <https://doi.org/10.1016/j.chb.2021.106715>
- 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. <https://doi.org/10.1016/j.chb.2014.09.012>
- 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. <https://doi.org/10.18637/jss.v020.i09>
- Maneesriwongul, W., & Dixon, J. K. (2004). Instrument translation process: A methods review. *Journal of Advanced Nursing*, *48*(2), 175–186. <https://doi.org/10.1111/j.1365-2648.2004.03185.x>
- McNeish, D. (2018). Thanks coefficient alpha, we'll take it from here. *Psychological Methods*, *23*(3), 412–433. <https://doi.org/10.1037/met0000144>
- Meyer, J. P., & Morin, A. J. (2016). A person-centered approach to commitment research: Theory, research, and methodology. *Journal of Organizational Behavior*, *37*(4), 584–612. <https://doi.org/10.1002/job.2085>

- Morin, A. J. S., & Marsh, H. W. (2015). Disentangling shape from level effects in person-centered analyses: An illustration based on university teachers' multidimensional profiles of effectiveness. *Structural Equation Modeling: A Multidisciplinary Journal*, 22(1), 39–59. <https://doi.org/10.1080/10705511.2014.919825>
- Mühling, A., Ruf, A., & Hubwieser, P. (2015). Design and first results of a psychometric test for measuring basic programming abilities. In A. Mühling, A. Ruf, & P. Hubwieser (Eds.), *Design and First Results of a Psychometric Test for Measuring Basic Programming Abilities* (pp. 2–10). ACM. <https://doi.org/10.1145/2818314.2818320>
- OECD. (2015). PISA 2012 Technical Report. OECD Publishing. <https://www.oecd.org/pisa/pisaproducts/PISA-2012-technical-report-final.pdf>
- OECD. (2017). PISA 2015 Technical Report. OECD Publishing. <http://www.oecd.org/pisa/sitedocument/PISA-2015-technical-report-final.pdf>
- Pellegrino, J. W., DiBello, L. V., & Goldman, S. R. (2016). A framework for conceptualizing and evaluating the validity of instructionally relevant assessments. *Educational Psychologist*, 51(1), 59–81. <https://doi.org/10.1080/00461520.2016.1145550>
- 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. <https://doi.org/10.1111/j.1745-3984.2006.00018.x>
- Polat, E., Hopcan, S., Kucuk, S., & Sisman, B. (2021). A comprehensive assessment of secondary school students' computational thinking skills. *British Journal of Educational Technology*, 52(5), 1965–1980. <https://doi.org/10.1111/bjjet.13092>
- Prenzel, M., Drechsel, B., & Kramer, K. (1998). Lernmotivation im kaufmännischen Unterricht: Die Sicht von Auszubildenden und Lehrkräften [Learning motivation in business education: the perspective of trainees and teachers]. *Zeitschrift für Berufs- Und Wirtschaftspädagogik*, special issue 14, 169–187
- Relkin, E., de Ruiter, L. E., & Bers, M. U. (2021). Learning to code and the acquisition of computational thinking by young children. *Computers & Education*, 169, 104222. <https://doi.org/10.1016/j.compedu.2021.104222>
- Repenning, A. (2006). Collaborative diffusion. Companion to the 21st ACM SIGPLAN Symposium on Object-Oriented Programming Systems, Languages, and Applications, 574–585. <https://doi.org/10.1145/1176617.1176630>
- Repenning, A. (2017). Moving beyond syntax: Lessons from 20 years of blocks programming in AgentSheets. *Journal of Visual Languages and Sentient Systems*, 3(1), 68–91. <https://doi.org/10.18293/VLSS2.017-010>
- Repenning, A. (2018). Scale or fail: Moving beyond self-selected computer science education in Switzerland. *Communications of the ACM*, 61(5), 40–42. <https://doi.org/10.1145/3199603>
- Repenning, A., Webb, D. C., Koh, K. H., Nickerson, H., Miller, S. B., Brand, C., Her Many Horses, I., Basawapatna, A., Gluck, F., Grover, R., Gutierrez, K., & Repenning, N. (2015). Scalable game design: A strategy to bring systemic computer science education to schools through game design and simulation creation. *ACM Transactions on Computing Education*, 15(2), 1–31. <https://doi.org/10.1145/2700517>
- Resnick, M., & Rusk, N. (2020). Coding at a crossroads. *Communications of the ACM*, 63(11), 120–127. <https://doi.org/10.1145/3375546>
- Richter, T., Naumann, J., & Horz, H. (2010). A revised version of the computer literacy inventory (INC-OBI-R). *Zeitschrift Für Pädagogische Psychologie*, 24(1), 23–37. <https://doi.org/10.1024/1010-0652/a000002>
- Robitzsch, A. (2020). Why ordinal variables can (almost) always be treated as continuous variables: Clarifying assumptions of robust continuous and ordinal factor analysis estimation methods. *Frontiers in Education*. <https://doi.org/10.3389/educ.2020.589965>
- Robitzsch, A., Kiefer, T., & Wu, M. (2020). Package 'TAM'. <https://cran.r-project.org/web/packages/TAM/TAM.pdf>
- Román-González, M. (2015). Computational thinking test: Design guidelines and content validation. <https://doi.org/10.13140/RG.2.1.4203.4329>
- 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. <https://doi.org/10.1016/j.chb.2016.08.047>
- 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). Springer. https://doi.org/10.1007/978-981-13-6528-7_6
- 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). ACM Press. <https://doi.org/10.1145/3290607.3312871>
- Rosenberg, J., Beymer, P., Anderson, D., van Lissa, C., & Schmidt, J. (2018). tidyLPA: An R package to easily carry out latent profile analysis (LPA) using open-source or commercial software. *Journal of Open Source Software*, 3(30), 978. <https://doi.org/10.21105/joss.00978>
- Rosman, T., Mayer, A.-K., & Krampen, G. (2015). Combining self-assessments and achievement tests in information literacy assessment: Empirical results and recommendations for practice. *Assessment & Evaluation in Higher Education*, 40(5), 740–754. <https://doi.org/10.1080/02602938.2014.950554>
- Rosseel, Y. (2012c). lavaan: An R package for structural equation modeling. *Journal of Statistical Software*, 48(2), 1–36. <https://doi.org/10.18637/jss.v048.i02>
- Salas, E., Reyes, D. L., & Woods, A. L. (2017). The assessment of team performance: observations and needs. In A. A. von Davier, M. Zhu, & P. C. Kyllonen (Eds.), *Innovative assessment of collaboration* (pp. 21–36). Springer. https://doi.org/10.1007/978-3-319-33261-1_2
- Sarstedt, M., Becker, J.-M., Ringle, C. M., & Schwaiger, M. (2011). Uncovering and treating unobserved heterogeneity with FIMIX-PLS: Which model selection criterion provides an appropriate number of segments? *Schmalenbach Business Review*, 63(1), 34–62. <https://doi.org/10.1007/BF03396886>
- Scherer, R., Howard, S. K., Tondeur, J., & Siddiq, F. (2021). Profiling teachers' readiness for online teaching and learning in higher education: Who's ready? *Computers in Human Behavior*, 118, 106675. <https://doi.org/10.1016/j.chb.2020.106675>
- Scherer, R., Siddiq, F., & Sánchez Viveros, B. (2019). The cognitive benefits of learning computer programming: A meta-analysis of transfer effects. *Journal of Educational Psychology*, 111(5), 764–792. <https://doi.org/10.1037/edu0000314>
- Scherer, R., Tondeur, J., & Siddiq, F. (2017). On the quest for validity: Testing the factor structure and measurement invariance of the technology-dimensions in the technological, pedagogical, and content knowledge (TPACK) model. *Computers & Education*, 112, 1–17. <https://doi.org/10.1016/j.compedu.2017.04.012>
- Seiter, L., & Foreman, B. (2013). Modeling the learning progressions of computational thinking of primary grade students. In B. Simon, A. Clear, & Q. Cutts (Eds.), *Proceedings of the ninth annual international ACM conference on International computing education research* (pp. 59–66). ACM. <https://doi.org/10.1145/2493394.2493403>
- Seufert, S., Guggemos, J., & Sailer, M. (2021). Technology-related knowledge, skills, and attitudes of pre- and in-service teachers: The current situation and emerging trends. *Computers in Human Behavior*, 115, 106552. <https://doi.org/10.1016/j.chb.2020.106552>
- Shute, V. J., Sun, C., & Asbell-Clarke, J. (2017). Demystifying computational thinking. *Educational Research Review*, 22, 142–158. <https://doi.org/10.1016/j.edurev.2017.09.003>
- Stadler, M., Herborn, K., Mustafić, M., & Greiff, S. (2020). The assessment of collaborative problem solving in PISA 2015: An investigation of the validity of the PISA 2015 CPS tasks. *Computers & Education*, 157, 103964. <https://doi.org/10.1016/j.compedu.2020.103964>
- Tang, X., Yin, Y., Lin, Q., Hadad, R., & Zhai, X. (2020). Assessing computational thinking: A systematic review of empirical studies. *Computers & Education*, 148, 103798. <https://doi.org/10.1016/j.compedu.2019.103798>
- Tondeur, J., Scherer, R., Baran, E., Siddiq, F., Valtonen, T., & Sointu, E. (2019). Teacher educators as gatekeepers: Preparing the next generation of teachers for technology integration in education. *British Journal of Educational Technology*, 50(3), 1189–1209. <https://doi.org/10.1111/bjet.12748>
- Ullrich, E., von Davier, M., & Pohl, S. (2020). A hierarchical latent response model for inferences about examinee engagement in terms of guessing and item-level non-response. *The British Journal of Mathematical and Statistical Psychology*, 73(Suppl 1), 83–112. <https://doi.org/10.1111/bmsp.12188>
- van de Schoot, R., Lugtig, P., & Hox, J. (2012). A checklist for testing measurement invariance. *European Journal of Developmental Psychology*, 9(4), 486–492. <https://doi.org/10.1080/17405629.2012.686740>
- van Laar, E., van Deursen, A. J., van Dijk, J. A., & de Haan, J. (2017). The relation between 21st-century skills and digital skills: A systematic literature review. *Computers in Human Behavior*, 72, 577–588. <https://doi.org/10.1016/j.chb.2017.03.010>
- Vanslambrouck, S., Zhu, C., Pynoo, B., Lombaerts, K., Tondeur, J., & Scherer, R. (2019). A latent profile analysis of adult students' online self-regulation in blended learning environments. *Computers in Human Behavior*, 99, 126–136. <https://doi.org/10.1016/j.chb.2019.05.021>

- 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. <https://doi.org/10.1007/s10639-015-9412-6>
- Werner, L., Denner, J., & Campe, S. (2012). The fairy performance assessment: measuring computational thinking in middle school. 43rd ACM Technical Symposium on Computer Science Education, 215–220. <https://doi.org/10.1145/2157136.2157200>
- Weintrop, D., Wise Rutstein, D., Bienkowski, M., & McGee, S. (2021). Assessing computational thinking: An overview of the field. *Computer Science Education*, 31(2), 113–116. <https://doi.org/10.1080/08993408.2021.1918380>
- Wilson, M. (2005). Constructing measures: An item response modeling approach. Psychology Press. <https://psycnet.apa.org/record/2004-22396-000>
- Willse, J. T. (2011). Mixture rasch models with joint maximum likelihood estimation. *Educational and Psychological Measurement*, 71(1), 5–19. <https://doi.org/10.1177/0013164410387335>
- Wing, J. M. (2006). Computational thinking. *Communications of the ACM*, 49(3), 33–35. <https://doi.org/10.1145/1118178.1118215>
- Wing, J. M. (2008). Computational thinking and thinking about computing. *Philosophical Transactions Series a, Mathematical, Physical, and Engineering Sciences*, 366(1881), 3717–3725. <https://doi.org/10.1098/rsta.2008.0118>
- Yadav, A., Hong, H., & Stephenson, C. (2016). Computational thinking for all: Pedagogical approaches to embedding 21st century problem solving in K-12 classrooms. *TechTrends*, 60(6), 565–568. <https://doi.org/10.1007/s11528-016-0087-7>
- Yadav, A., Mayfield, C., Zhou, N., Hambrusch, S., & Korb, J. T. (2014). Computational thinking in elementary and secondary teacher education. *ACM Transactions on Computing Education*, 14(1), 1–16. <https://doi.org/10.1145/2576872>
- Zhao, W., & Shute, V. J. (2019). Can playing a video game foster computational thinking skills? *Computers & Education*, 141, 103633. <https://doi.org/10.1016/j.compedu.2019.103633>

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.