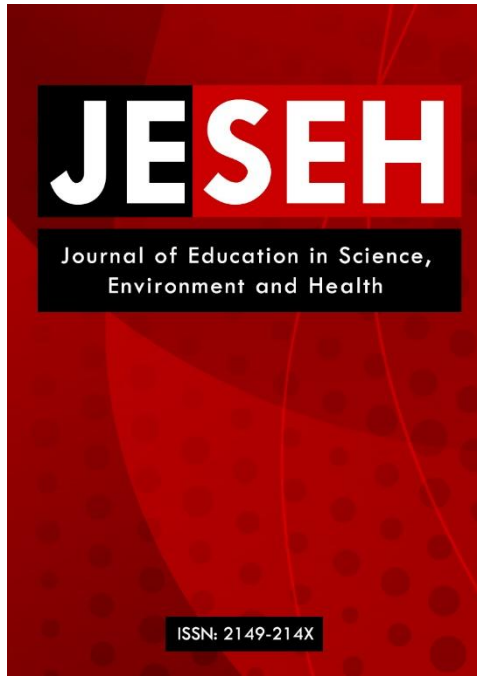




ISSN: 2149-214X

**Journal of Education in Science,
Environment and Health**

www.jeseh.net

A Reliability Generalization Meta-Analysis of the Computational Thinking Test (CTt): Evidence from STEM Education

Seyma Irmak¹, Kaan Bati²

¹Amasya University

²Hacettepe University

To cite this article:

Irmak, S., & Bati, K. (2026). A reliability generalization meta-analysis of the computational thinking test (CTt): Evidence from STEM education. *Journal of Education in Science, Environment and Health (JESEH)*, 12(4), 371-387. <https://doi.org/10.55549/jeseh.928>

This article may be used for research, teaching, and private study purposes.

Any substantial or systematic reproduction, redistribution, reselling, loan, sub-licensing, systematic supply, or distribution in any form to anyone is expressly forbidden.

Authors alone are responsible for the contents of their articles. The journal owns the copyright of the articles.

The publisher shall not be liable for any loss, actions, claims, proceedings, demand, or costs or damages whatsoever or howsoever caused arising directly or indirectly in connection with or arising out of the use of the research material.

A Reliability Generalization Meta-Analysis of the Computational Thinking Test (CTt): Evidence from STEM Education

Seyma Irmak, Kaan Bati

Article Info	Abstract
Article History Published: 01 October 2026 Received: 21 June 2026 Accepted: 22 September 2026 Keywords 21st century abilities, Computational thinking STEM education	This study performs a Reliability Generalization (RG) meta-analysis — a quantitative synthesis of reliability coefficients reported across independent studies — to synthesize internal consistency estimates of the CTt and identify sources of variability. Computational Thinking Test (CTt; Román-González et al., 2017) is one of the most widely adopted assessment tools in the field. The random-effects model revealed a pooled Cronbach's alpha coefficient of 0.77 (95% confidence interval [0.75, 0.79]), indicating robust reliability consistent with the original validation study. Despite this acceptable average, substantial heterogeneity was observed ($I^2 = 81\%$). Moderator analyses identified school level as a significant predictor of reliability ($p < .01$). Specifically, reliability estimates were significantly lower for primary school samples compared to higher education students, suggesting that cognitive maturity and test-taking skills may influence measurement error in younger populations. Conversely, language, country, and sample size did not significantly moderate reliability, supporting the instrument's cross-cultural applicability.

Introduction

Originally introduced from a conceptual perspective by Papert (1980), computational thinking (CT) was later redefined and popularized by Wing (2006), who argued that, alongside reading, writing, and arithmetic, CT should be regarded as a fundamental analytical skill that every child ought to acquire. This perspective has contributed to a growing consensus that, in order to prepare children for participation in a digital society, CT should be fostered from the early years together with creativity, problem solving, and digital competence, thereby accelerating both policy initiatives and scholarly work in this field (Barr et al., 2011; Bocconi et al., 2016; Yang et al., 2023). Similarly, Vee (2017) emphasized that technological developments possess transformative potential for education and give rise to lifelong competencies such as coding, programming, and CT in the digital age. Consequently, CT has increasingly been positioned as a powerful vehicle for fostering 21st-century skills (Falloon, 2024; Shute et al., 2017). As a cognitive process that involves reasoning and higher-order thinking, CT is widely assumed to support multiple aspects of children's cognitive development, including critical thinking, creativity, and problem solving, which underscores the importance of introducing CT from an early age (Bati, 2022). Within the specific context of STEM education, this broader emphasis on CT has been formalized through national curricular standards. Notably, the Next Generation Science Standards (NGSS Lead States, 2013) designate "Using Mathematics and Computational Thinking" as one of the eight core Science and Engineering Practices, explicitly requiring K–12 students to construct algorithms, analyze large data sets, and develop simulations of natural and designed systems. Complementing this policy-level recognition, Weintrop et al. (2016) operationalized CT specifically for mathematics and science classrooms through a four-category taxonomy comprising data practices, modeling and simulation practices, computational problem-solving practices, and systems thinking practices. Taken together, these frameworks position CT not merely as a generic 21st-century competency but as an integral component of STEM disciplinary practice, thereby reinforcing the need for psychometrically sound instruments capable of capturing CT development within STEM learning contexts. Beyond this standards-based recognition, empirical research continues to examine how CT can be cultivated within STEM learning environments. In a quasi-experimental study with Nigerian undergraduates, AI-supported problem-based learning produced significantly greater gains in CT and academic achievement than conventional problem-based instruction (Ayanwale & Omeh, 2026). Similarly, structural equation modeling has revealed strong positive associations between STEM attitudes, CT, and creative thinking among secondary school students (Suherman et al., 2025), while robotics-integrated game design has been shown to yield higher CT ratings when students engage in holistic, goal-directed development rather than tutorial-based instruction (Leonard et al., 2016). A recent systematic review further identified project-based learning as a central pedagogical strategy for developing CT within STEM higher education (Tariq et al., 2025). This sustained empirical interest in fostering CT within STEM contexts further underscores the need for psychometrically robust instruments capable of reliably capturing such gains.

Despite this growing interest, CT remains a relatively new construct that is still in its formative stage. As noted by Tsarava et al. (2022), there is still no clear, empirically grounded, and widely accepted definition of CT in the literature. Likewise, Lee, et al. (2023) argue that little is currently known about how CT can be effectively supported in the early years. These gaps highlight the urgent need to clarify the conceptual boundaries of CT, to design and evaluate instructional approaches and learning materials that foster CT, and, critically, to develop assessment frameworks and instruments that can measure CT skills and related curricula in a valid and reliable manner. This study performs a Reliability Generalization (RG) meta-analysis to synthesize internal consistency estimates of the CTt (Román-González, et al., 2017) and identify sources of variability.

Assessing CT Skills

Among the instruments developed to assess CT, the Computational Thinking Test (CTt; Román-González et al., 2017) is one of the most widely adopted assessment tools in the field. Grounded theoretically in the frameworks proposed by Brennan and Resnick (2012) and Wing (2006), the CTt comprises 28 multiple-choice items presented through a block-based visual programming interface, requiring examinees to solve tasks involving basic and compound sequences, simple and complex loops, and simple and complex conditionals. These item types closely parallel the computational problem-solving and data practices identified by Weintrop et al. (2016) as core to STEM disciplinary reasoning, situating the CTt as a measure with direct relevance to STEM learning contexts. In its original validation study, Román-González et al. (2017) reported Cronbach's alpha coefficients of .72, .76, and .82 across different educational levels. Criterion-related validity was further supported by statistically significant correlations between CTt scores and independent measures of spatial ability ($r = .44$), reasoning ability ($r = .44$), and problem-solving ability ($r = .67$), a pattern interpreted as consistent with the Cattell-Horn-Carroll (CHC) model of intelligence. Owing to this initial psychometric support and its free availability, the CTt has since been adopted across a wide range of educational and cultural contexts. Despite the growing number of instruments developed in recent years to assess CT skills, many of these tools fail to meet essential measurement criteria (Hubwieser & Mühling, 2014; Bellettini et al., 2015; Gane et al., 2021; Parker et al., 2021). For instance, the Bebras Challenge is sometimes employed as an assessment of CT skills; however, it has undergone only limited psychometric validation, which raises concerns about its use for rigorous research purposes (Hubwieser & Mühling, 2014; Bellettini et al., 2015). Similarly, the assessment proposed by Gane et al. (2021) requires manual scoring and multiple raters, thereby constraining its scalability and limiting its usability by other researchers and practitioners. The instrument developed by Parker et al. (2021), which combines block-based and Bebras-style items, was piloted with only 57 fourth-grade students, restricting the generalizability of the findings. Likewise, the assessment designed by Chen et al. (2017) for fifth-grade students appears to be highly dependent on a robotics programming context, includes open-ended items, and was administered to only 37 students. Consequently, these instruments share common methodological limitations, including small sample sizes, contextual dependency, and challenges in administration and scoring.

In contrast, several instruments have undergone psychometric validation at the elementary school level and meet the aforementioned criteria. These include TechCheck for Grades 1–2 (ages 6–8; Relkin et al., 2020), TechCheck-K for kindergarten (ages 4–6; Relkin & Bers, 2021), the Beginners Computational Thinking Test (BCTt) for kindergarten through Grade 6 (ages 5–10; Zapata-Cáceres et al., 2020), the competent Computational Thinking test (cCTt) for Grades 3–4 (ages 7–9; El-Hamamsy et al., 2022), the Computational Thinking Assessment for Chinese Elementary Students (CTA-CES) for Grades 3–6 (ages 9–12; Li et al., 2021), and the CT concepts test for Grades 3–5 developed by Kong and Lai (2022). Moreover, all of these tools are relatively recent and adopt an unplugged approach using multiple-choice items to assess elementary students' CT abilities. Considerable overlap also exists in their target age ranges. As a result, researchers and practitioners must not only identify instruments that best align with the learning objectives of their interventions, but also understand the validity boundaries of these tools in order to make informed methodological decisions. Unfortunately, these instruments have rarely been systematically compared to determine which is most appropriate for specific age groups. To the best of our knowledge, only TechCheck and TechCheck-K have been compared to examine suitability for kindergarten students, with results suggesting that TechCheck-K is more appropriate for kindergarten, whereas TechCheck is better suited for first- and second-grade students (Relkin & Bers, 2021).

Research Questions

1. What is the average reliability coefficient of the CTt scores across empirical studies?
2. To what extent do study characteristics (e.g., publication year, test language) and sample characteristics (e.g., educational level, gender distribution) moderate the variability in CTt score reliability?

Method

In contemporary psychometric theory, reliability is not considered an immutable attribute of a measurement instrument itself—such as the Computational Thinking Test (CTt)—but rather a property of the scores obtained from a specific group of examinees (Thompson & Vacha-Haase, 2000; Messick, 1995). Consequently, reliability estimates can fluctuate significantly depending on sample heterogeneity and administration contexts. This context-dependency arises largely because internal-consistency estimates such as Cronbach's alpha are a direct function of the interitem covariance structure observed within a given sample: more heterogeneous groups of examinees tend to yield higher alpha values, whereas more homogeneous or restricted-range samples tend to yield lower ones, independent of any change in the underlying test itself (Sijtsma, 2009). To address this, the present study employs the Reliability Generalization (RG) meta-analytic technique originally proposed by Vacha-Haase (1998). Unlike a narrative review, which merely catalogues reliability values reported across studies, RG treats reliability coefficients themselves as the unit of analysis: coefficients are statistically weighted by their precision, submitted to a variance-stabilizing transformation to satisfy distributional assumptions, and then pooled to yield a single, weighted estimate of reliability alongside a formal test of the extent to which observed differences across studies exceed what would be expected from sampling error alone. Furthermore, this method allows for the assessment of variability among these estimates and the identification of potential moderator variables—such as educational level or test language—that may influence the consistency of test scores (Botella et al., 2010). This approach provides a comprehensive understanding of the conditions under which the CTt functions most effectively. Throughout the planning, execution, and reporting phases of this study, we strictly adhered to the guidelines of the Reliability Generalization Meta-Analysis (REGEMA) checklist established by Sánchez-Meca et al. (2021).

Selection Criteria and Search Strategies

In this meta-analysis, we encompassed all iterations and validated adaptations of the Computational Thinking Test (CTt), originally developed by Román-González et al. (2017), found in the empirical literature. To ensure methodological rigor, we defined precise eligibility standards for study inclusion and exclusion. A primary inclusion criterion was that studies must have utilized this specific instrument and explicitly reported internal consistency reliability in the form of Cronbach's alpha. The decision to prioritize Cronbach's alpha as the focal metric was driven by its ubiquity in educational research and its status as the most widely recognized indicator of scale reliability. It is worth noting that Kuder-Richardson Formula 20 (KR-20) is mathematically equivalent to Cronbach's alpha for dichotomously scored items such as those comprising the CTt (Kuder & Richardson, 1937); accordingly, KR-20 coefficients would, in principle, have been eligible for inclusion on the same basis as Cronbach's alpha. However, none of the 26 studies retained in the final dataset reported KR-20 as their reliability metric. Consequently, restricting our analysis to this metric ensured the most robust and comparable dataset for synthesizing reliability estimates across diverse research contexts.

Our electronic search strategy was designed to achieve a strategic balance between methodological quality and comprehensive coverage. To this end, we utilized the Web of Science (WoS) Core Collection as our primary source, predicated on its reputation as the premier global index of high-impact scientific research. This ensured that the core of our dataset was derived from publications meeting rigorous methodological standards. Complementing this, we integrated Google Scholar into our search process. This addition served to mitigate potential publication bias and to identify relevant studies that might not yet be indexed in WoS or were published in journals with different indexing scopes. Despite the breadth of Google Scholar, we maintained strict quality control by prioritizing peer-reviewed journal articles across both platforms. This focus ensured that the reliability estimates synthesized in this study were derived from sources that have undergone a peer-review process, thereby maximizing the methodological quality of the included data while ensuring no relevant study was overlooked. Furthermore, we restricted the document type to journal articles, excluding conference proceedings and gray literature (e.g., dissertations). This restriction was applied to ensure that the reliability estimates synthesized in this study were derived from publications that have undergone a rigorous peer-review process, thereby maximizing the methodological quality of the included data. The temporal scope of our search spanned from 2017, coinciding with the publication of the seminal work on the CTt by Román-González et al. (2017), to December 23, 2024. Regarding language criteria, the review was restricted to studies published in English. This limitation was applied to maintain linguistic consistency in data extraction and reflects the status of English as the dominant lingua franca in the field of CT research. Finally, no geographical or cultural restrictions were imposed, allowing for a global assessment of the test's psychometric performance across different populations.

Our exclusion criteria were meticulously defined to ensure the integrity of the dataset. Firstly, despite utilizing institutional proxy access to retrieve full texts, studies that remained inaccessible (e.g., due to strict paywalls or unavailable digital archives) were inevitably excluded. More critically, our analysis necessitated a distinction between studies that provided empirical evidence of reliability and those that neglected to do so. Following the framework established in previous reliability generalization research (e.g., Vacha-Haase, 1998), we categorized studies into induction by report and induction by omission. Induction by report refers to studies that explicitly calculated and presented the Cronbach's alpha coefficient for the CTt data collected in their specific sample. Conversely, induction by omission pertains to studies that utilized the CTt but failed to report reliability coefficients for their own data, often relying instead on the psychometric properties reported in the original test manual. We calculated the rates for both categories to transparently depict the reporting practices within the CTt literature. Specifically, studies failing to provide a Cronbach's alpha value for the overall test score were excluded from the meta-analysis, as our primary focus was on the reliability of the total score. Additionally, studies where the abstract was in English but the main text was in a different language were also removed during the full-text screening phase.

In crafting our search strategies for electronic sources, we meticulously identified key terms essential across all searches, aligning strictly with our eligibility criteria. The primary terms selected explicitly referenced the various nomenclatures of the instrument, such as "Computational Thinking Test," "CT test," or its abbreviation "CTt," ensuring comprehensive coverage of studies employing this specific tool.

For each database, we tailored search strings in accordance with their distinct search logics and functionalities to balance sensitivity and precision. For the WoS Core Collection, we utilized the "Advanced Search" feature to execute a broad, highly sensitive query using the ALL field tag: (ALL=("computational thinking test" OR "computational thinking skills test" OR "CT test" OR "CTt")). A distinct aspect of our WoS strategy was the deliberate omission of the developers' names (i.e., Román-González) from the search code. This approach was intended to capture studies that utilized the test but might not have cited the authors in the indexed metadata, thereby maximizing inclusivity within peer-reviewed literature.

Conversely, for Google Scholar, we adapted our strategy to accommodate its extensive full-text indexing mechanism. Given that searching for the acronym "CTt" alone would generate an unmanageable volume of irrelevant results (e.g., medical imaging studies), we incorporated the primary developer's surname to optimize relevance. Thus, the search string ("CTt" AND "Román-González") was employed. This tailored approach allowed us to filter out unrelated disciplines while still capturing relevant peer-reviewed journal articles that were not indexed within the WoS Core Collection. Across both platforms, the search spanned from 2017 to December 23, 2024. By the conclusion of the process, our electronic search had identified a total of 2,197 papers from WoS and 402 from Google Scholar.

Data Extraction

Following the retrieval of search results, we organized the publications chronologically and stored them for systematic review. Subsequently, we engaged in a preliminary examination of the selected articles to identify potential variables that could serve as moderators. Our objective was to pinpoint specific study and sample characteristics that might influence the reliability outcomes. From each eligible article, we extracted essential data related to the CTt, including the reported Cronbach's alpha coefficient, sample size (N), the number of items in the administered test, the country where the study was conducted, the educational level of the participants, the language of the test administration, the percentage of female participants, the study setting (urban/rural), and the standard deviation (SD).

We considered sample size as a moderator based on the premise that larger samples typically yield more stable and generalizable reliability estimates, minimizing sampling error. We included the number of items as a categorical moderator (Original 28-items vs. Modified forms) because test length is theoretically linked to internal consistency; scales with more items generally produce higher reliability due to increased variance. Additionally, we took into account the country of the study to address potential cross-cultural differences in educational contexts, which might influence how CT is understood and measured.

We coded the educational level (Primary, Lower Secondary, Upper Secondary, Higher Education) after recognizing that cognitive maturity and developmental stages can impact students' engagement with CTt tasks, thereby affecting reliability estimates. We also considered the language of the test as a critical moderator. Even though the meta-analysis was restricted to English-language publications, the test itself was often administered in

various local languages. Translation and cultural adaptation processes can alter item nuances and response patterns, potentially influencing the internal consistency of the CTt in different linguistic settings.

Regarding gender, we considered the percentage of female participants because gender differences in CT interest are well-documented in the literature. Variations in the proportion of female participants could influence the homogeneity of the sample, potentially affecting scale reliability. We also incorporated the standard deviation of test scores as a moderator to empirically understand the relationship between score variability and reliability coefficients. According to classical test theory, higher variability (SD) in test scores is often associated with higher reliability estimates.

Lastly, we coded the study setting as a dichotomous variable (Urban vs. Rural) rather than a continuous percentage, as most studies were conducted exclusively in one setting. This allowed us to assess whether differences in educational resources and infrastructure between urban and rural environments impact the consistency of CTt scores. By considering these moderators, we aimed to conduct a comprehensive analysis that accounted for various factors influencing the reliability of the CTt, thereby providing a thorough understanding of the instrument's performance across different settings and conditions.

Given the nascent and rapidly evolving nature of CT education, we acknowledged that certain contextual variables might theoretically influence the reliability of CTt scores. Specifically, factors such as students' prior coding experience (e.g., knowledge of Scratch or Python), the availability of computational resources in schools (e.g., computer labs vs. tablets), or the specific teaching curriculum followed were seldom reported with sufficient consistency across the included studies. Consequently, despite their potential relevance, we were unable to include these variables in our set of moderators. Furthermore, we deliberately excluded demographic variables such as ethnicity or socio-economic status, as these were not directly reported in the majority of the primary studies, and inferring them would introduce significant measurement error. Details of the recorded variables and their coding scheme are outlined in Table 1.

Table 1. Coding system for identified moderators

Moderator	Type	Coding Method
Sample size	Continuous	Each study was assigned a unique numerical code
Percentage of female participants in the sample	Continuous	Each study was assigned a unique numerical code
SD of test score	Continuous	Each study was assigned a unique numerical code
Publication year	Continuous	Each study was assigned a unique numerical code
Item number	Categorical	1 = 28 items (original form); 2 = 28 items (modified forms)
Country	Categorical	1 = Hungary; 2 = Spain; 3 = Netherlands; 4 = Turkey; 5 = Switzerland; 6 = Singapore; 7 = Germany; 8 = Nigeria; 9 = Finland; 10 = China; 11 = Taiwan; 12 = Greece
School level	Categorical	1 = Primary education; 2 = Lower secondary education; 3 = Upper secondary education; 4 = Higher education
Area type	Categorical	1 = Urban; 2 = Rural
Language of the test administration	Categorical	1 = English; 2 = Hungarian; 3 = Spanish; 4 = Turkish; 5 = German; 6 = Greek; 7 = Chinese; 8 = N/A

All extracted data were systematically recorded in a structured spreadsheet designed to capture the maximum number of relevant moderators. To mitigate potential coding errors and minimize selection bias, a rigorous reliability check was implemented. The entire selection and coding process were executed in two distinct phases. In the first phase (Study Selection), the primary researcher applied the predefined inclusion/exclusion criteria to screen the retrieved records. To assess consistency, a random sample of 20% of the initial records was re-screened by the same researcher after a two-week interval. The intra-rater agreement for this phase was calculated using Cohen's Kappa, yielding a value of 0.85, indicating substantial consistency.

In the second phase (Variable Coding), the selected studies were coded for both categorical and continuous moderators. Similarly, an intra-rater reliability check was performed on a random subset of the included studies after a temporal delay. For this phase, the agreement was exceptionally high (Kappa = 0.96), demonstrating that the extraction of objective data (e.g., sample size, reliability coefficients) was highly accurate and reproducible. Any minor discrepancies identified during the re-coding process were resolved by revisiting the original full-text articles.

Statistical Analysis

The statistical analysis was rigorously structured to synthesize the reliability estimates of the CTt while addressing the specific distributional properties of the data. As noted in the psychometric literature, Cronbach's alpha coefficients typically exhibit negative skewness due to their bounded nature and the high inter-item correlations inherent in internal consistency measures (Beretvas & Pastor, 2003). To correct this non-normal distribution and satisfy the assumptions required for meta-analytic integration, Bonett's transformation was applied to the raw reliability coefficients using the following formula: $Li = \ln(1 - \alpha_i)$. This transformation is empirically preferred over Fisher's z transformation for Cronbach's alpha coefficients because it provides more accurate variance stabilization (Bonett, 2002). Following the meta-analytic procedures, pooled reliability estimates and their confidence intervals were back-transformed to the original metric using the inverse formula: $\alpha_i = 1 - e^{Li}$. This back-transformation facilitated the interpretation of the results in terms of the original reliability coefficient. Given the diversity of the included studies with respect to educational levels, cultural contexts, and test adaptations, it was assumed that the true reliability of the CTt varied across populations. Therefore, a random-effects model was employed rather than a fixed-effects model, allowing for unconditional inferences about the broader universe of studies (Borenstein et al., 2010). The between-study variance component (τ^2) was estimated using the restricted maximum likelihood (REML) method, which provides less biased estimates of τ^2 than the DerSimonian-Laird method, particularly when the number of studies is relatively small (Veroniki et al., 2016). Within this framework, individual studies were weighted using the inverse variance method, which assigns greater weight to studies with higher precision and thereby optimizes the accuracy of the pooled estimate (Marín-Martínez & Sánchez-Meca, 2010). To evaluate the consistency of reliability estimates across studies, Cochran's Q statistic was computed to test the null hypothesis of homogeneity. In addition, the I-squared (I^2) index was calculated to quantify the proportion of observed variance attributable to real heterogeneity rather than sampling error (Higgins et al., 2003). An I^2 value greater than 75% was interpreted as indicating substantial heterogeneity (Higgins et al., 2003), thereby justifying the inclusion of moderator analyses. After confirming the presence of heterogeneity, potential sources of variance were examined using mixed-effects models. Analysis of Variance (ANOVA) was conducted for categorical moderators, including country, school level, test version, area type, and language of test administration. Meta-regression analyses were performed for continuous moderators, namely sample size, publication year, percentage of female participants, and test score standard deviation. Each moderator was tested individually rather than jointly, given the limited number of independent samples ($k = 26$) relative to the number of candidate moderators. Because the Country (13 categories) and Language (7 categories) moderators contained several categories represented by a single study, these variables were additionally re-analyzed using theoretically motivated, more balanced groupings—Country as a four-level region variable (Europe, Asia, Americas, Africa) and as a binary Europe-versus-rest-of-world contrast, and Language as a binary Spanish-versus-non-Spanish contrast—to evaluate whether the original fine-grained categorization affected the conclusions. Because not all included studies reported the percentage of female participants, the standard deviation of test scores, or the urban/rural study setting, these three moderator analyses were conducted using an available-case (pairwise deletion) approach, in which only studies reporting the relevant variable were included in the corresponding model; the resulting sample size for each moderator is reported alongside its results in Tables 3 and 4.

To assess the robustness of the findings and identify potential publication bias, multiple complementary methods were employed. Rosenthal's Fail-safe N (Rosenthal, 1979) was used to estimate the number of non-significant studies required to reduce the overall effect to non-significance. Begg and Mazumdar's rank correlation test (Begg & Mazumdar, 1994) and Egger's regression test (Egger et al., 1997) were applied to statistically evaluate funnel plot asymmetry. In addition, the Duval and Tweedie trim-and-fill method (Duval & Tweedie, 2000) was used to estimate and adjust for the potential influence of missing studies. To evaluate the robustness of the pooled reliability estimate to the inclusion of small-sample studies, sensitivity analyses were conducted by re-estimating the model after excluding studies with $N < 30$ and, separately, $N < 50$. A leave-one-out analysis was additionally performed to assess the influence of any single study on the pooled estimate. All statistical analyses were conducted using the R statistical computing environment (R Core Team, 2024). The metafor package (Viechtbauer, 2010) was primarily used for meta-analytic computations, supported by the dplyr (Wickham et al., 2022), readxl (Wickham & Bryan, 2023), tibble (Müller & Wickham, 2023), tidyr (Wickham et al., 2024), rstatix (Kassambara, 2023), writexl (Ooms, 2024), and ggplot2 (Wickham, 2016) packages for data management, statistical support, and visualization.

Results

Our comprehensive search strategy, encompassing the WoS Core Collection and a supplementary search on Google Scholar, initially identified a total of 2,599 records. Given the comprehensive indexing of Google Scholar,

a significant overlap with the WoS results was anticipated and observed; 186 records from the Google Scholar search were identified as duplicates of the WoS entries and were removed, leaving 2,413 unique records for screening. Following the screening of titles and abstracts, 2,271 records were excluded: 2,075 were deemed irrelevant or did not utilize the CTt, 121 were systematic reviews, meta-analyses, or other secondary research types, and 75 were not published in English. This left 142 reports to be sought for full-text retrieval. Of these, 26 studies (5 from WoS and 21 from Google Scholar) could not be retrieved due to paywalls or a lack of institutional access, leaving 116 reports assessed for eligibility. The induction by report rate, which indicates the proportion of accessible studies utilizing the CTt that explicitly reported the Cronbach's alpha coefficient for their sample, was 22.41% (26 out of 116 studies examined in full text). Conversely, the induction by omission rate, reflecting the proportion of studies that utilized the CTt but failed to report the reliability coefficient for their specific data, was 77.59% (90 out of 116 studies). This high omission rate, exceeding the commonly observed thresholds, indicates that a vast majority of researchers utilizing the CTt rely on psychometric evidence from the original validation study rather than recalculating reliability for their own samples. While this suggests a potential reporting bias, it is reflective of a broader trend in the educational technology literature and does not necessarily invalidate the findings derived from the reporting studies. Therefore, we proceeded with the analysis, confident that the included sample provides a robust representation of the available reliability data. The PRISMA flowchart illustrating the detailed study selection process is presented in Figure 1.

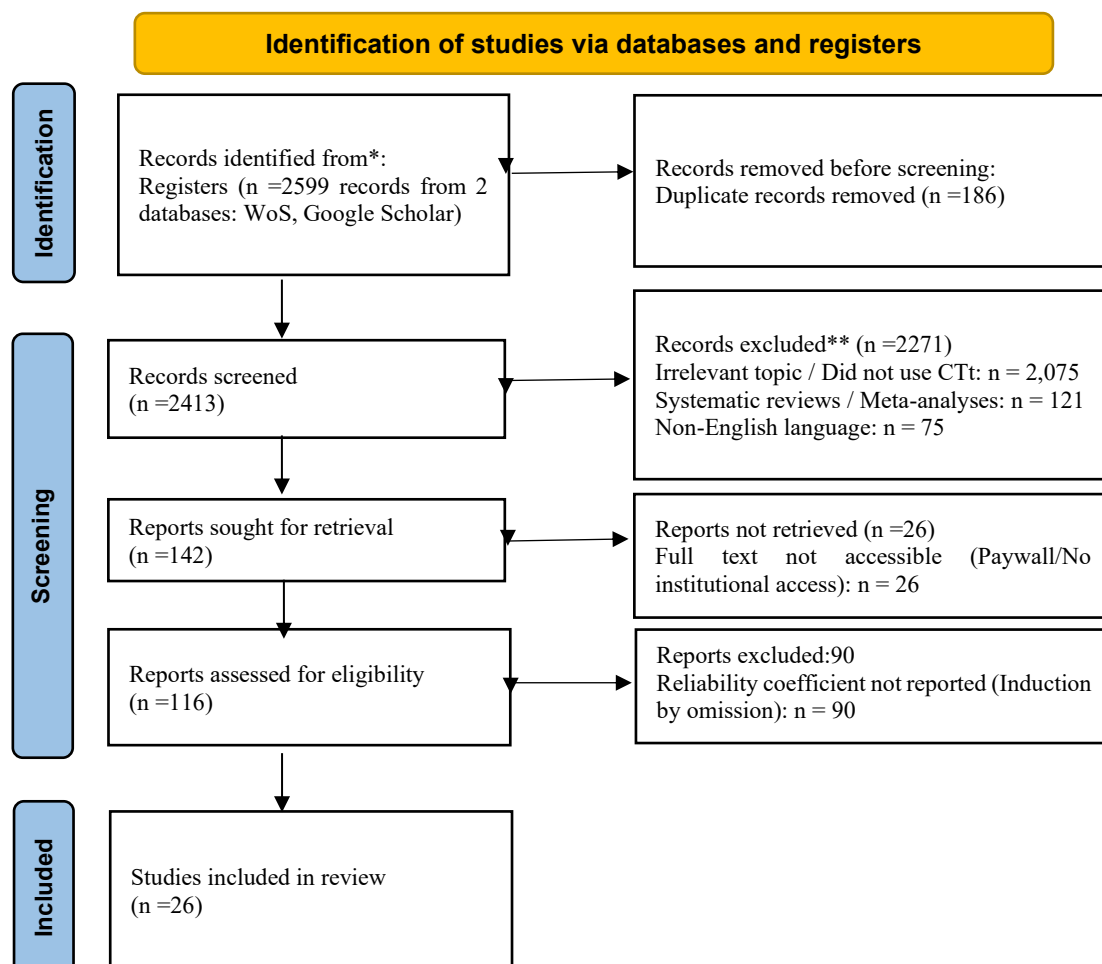


Figure 1. PRISMA flowchart depicting study selection process for CTt reliability meta-analysis

Following these rigorous exclusions, a total of 26 independent samples, drawn from 24 unique publications, met all inclusion criteria and provided the necessary data for the meta-analysis. In one case—Román-González et al. (2017)—a single publication reported separate, non-overlapping reliability coefficients for distinct educational-level subgroups (three subgroups); each subgroup was treated as an independent sample in the meta-analysis, consistent with standard practice in reliability generalization studies (Vacha-Haase, 1998), given that no participants were shared across subgroups. These 24 publications, published between 2017 and 2025, encompassed a wide range of sample sizes, with the smallest sample including 21 participants and the largest comprising 2,217 participants. Geographically, the included studies represented a diverse set of contexts, predominantly from Europe and Asia. The distribution of studies by country was as follows: Spain (30.77%),

Netherlands (15.38%), China (7.69%), Greece (7.69%), Switzerland (7.69%), with single studies representing Hungary, Turkey, Singapore, Germany, Nigeria, Taiwan, Chile, and Finland (3.85% each). Regarding the educational level of the sampled groups, the CTt was most frequently administered in primary education (38.46%), followed by higher education (26.92%), lower secondary education (15.38%), and upper secondary education (15.38%); one additional sample (Moreno-León et al., 2024; 3.85%) spanned a combined primary-to-lower-secondary population and could not be classified into a single educational level (see Section 3.3). In terms of instrument structure, the vast majority of studies utilized the original 28-item version of the CTt (80.77%), while a smaller proportion employed modified versions (e.g., shortened or adapted forms) (19.23%). The linguistic diversity of the test administration was also noted. While the reporting language of the articles was English, the CTt itself was administered in various languages: Spanish (30.77%), German (11.54%), Greek (7.69%), and single instances of English, Chinese, Hungarian, and Turkish (3.85% each). In 9 studies (34.62%), the specific language of the test administration was not explicitly stated. Table 2 provides a detailed summary of the studies included in this meta-analysis.

Table 2. Summary of 26 reliability coefficients

Authors	Language	Country	School Level	Sample Size	Item Number	Cronbach's Alpha
Aranyi et al., 2024	Hungarian	Hungary	Higher Education	97	28	.79
Martín et al., 2024	Spanish	Spain	Lower Secondary	85	28	.76
Villalustre & Cueli, 2023	Spanish	Spain	Higher Education	164	28	.70
Fanchamps et al., 2024	N/A	Netherlands	Primary	35	28	.68
Bilgic & Dogusoy, 2023	Turkish	Turkey	Primary	62	28	.79
Guggemos et al., 2022	German	Switzerland	Upper Secondary	202	28	.78
Fanchamps et al., 2021	N/A	Netherlands	Primary	30	28	.67
Guggemos, 2021	German	Switzerland	Upper Secondary	202	26	.74
Chan et al., 2020	N/A	Singapore	Upper Secondary	153	26	.77
Sarava et al., 2019	German	Germany	Primary	31	21	.76
Moreno-León et al., 2024	Spanish	Spain	Mixed (K4–K8)	2217	28	.791
Yusuf et al., 2024	N/A	Nigeria	Higher Education	35	28	.84
Peracaula-Boscha & González-Martínez, 2024	Spanish	Spain	Higher Education	37	28	.75
González-Martínez et al., 2024	Spanish	Spain	Higher Education	71	28	.83
De la Hoz Serrano et al., 2024	English	Finland	Higher Education	116	14	.79
Cheng et al., 2024	N/A	China	Primary	857	28	.72
Fanchamps et al., 2024	N/A	Netherlands	Lower Secondary	21	28	.82
Chen & Huang, 2024	N/A	Taiwan	Primary	48	28	.79
Tikva & Tambouris, 2023	Greek	Greece	Lower Secondary	57	28	.80
Fanchamps et al., 2023	N/A	Netherlands	Primary	156	28	.67
González-Pizarro et al., 2023	N/A	Chile	Higher Education	549	32	.87
Kakavas & Ugolini, 2022	Greek	Greece	Primary	85	28	.70
Wang et al., 2022	Chinese	China	Primary	84	28	.81
Román-González et al., 2017	Spanish	Spain	Primary	176	28	.72
Román-González et al., 2017	Spanish	Spain	Lower Secondary	735	28	.76
Román-González et al., 2017	Spanish	Spain	Upper Secondary	340	28	.82

Publication Bias

Our statistical assessment did not detect significant evidence of publication bias in the dataset. First, the Begg and Mazumdar rank correlation test revealed that the correlation between the reliability coefficients and their variances was negative and statistically non-significant (Kendall's $\tau = -0.0772$, $p = 0.581$). Corroborating this finding, Egger's regression test for funnel plot asymmetry also yielded a z-value that was not statistically significant ($z = 0.2975$, $p = 0.766$). The non-significant probability values obtained from these two complementary tests provide no evidence of systematic bias in the retrieval of studies, although the power of these tests to detect bias is limited when the number of included samples is relatively small ($k = 26$).

To further test the robustness of the findings against the "file-drawer problem," we calculated Rosenthal's Fail-safe N. The analysis produced an exceptionally high value of 44,044, indicating that more than forty-four thousand unpublished studies with null results would be required to reduce the current meta-analytic mean to a non-significant level. Finally, employing the Duval and Tweedie trim-and-fill method resulted in zero (0) imputed missing studies, confirming the symmetry of the funnel plot (see Fig. 2).

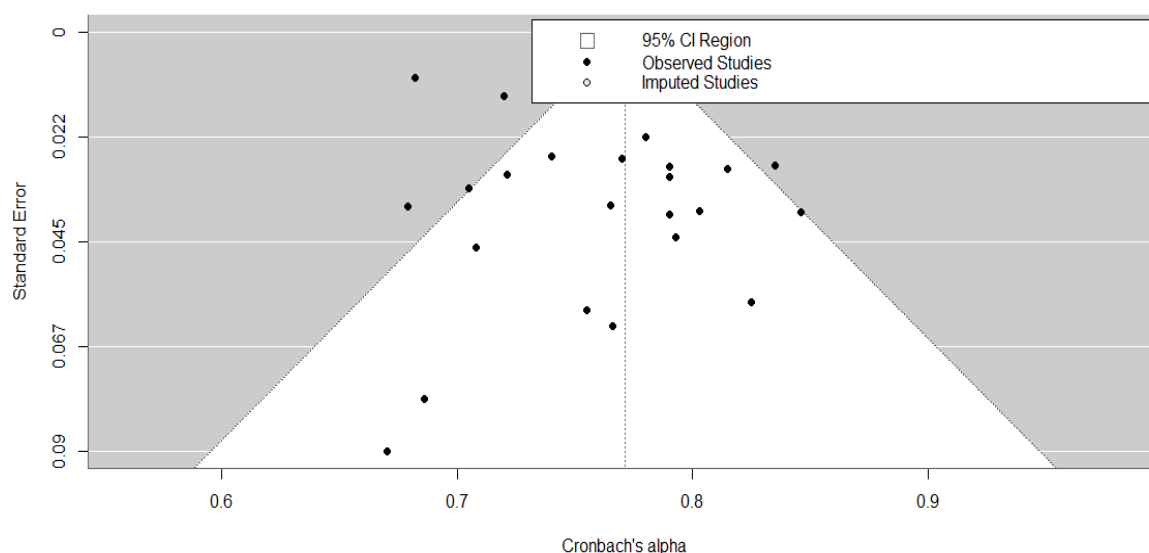


Figure 2. Funnel plot displaying Bonett's transformed alpha coefficient for the CTt total scale

Since no missing studies were identified, the adjusted mean reliability coefficient remained identical to the observed pooled estimate ($\alpha = .773$). Based on these convergent findings, publication bias does not appear to be a substantial threat to the overall reliability estimation of the CTt in this study.

Mean Reliability and Heterogeneity

The meta-analytic synthesis of the 26 independent samples yielded a pooled Cronbach's alpha coefficient of .77 (95% CI [.75, .79], $SE = .011$, $z = 68.71$, $p < .0001$), indicating that, on average, the CTt demonstrates an acceptable level of internal consistency reliability. Given the presence of variability across studies, this pooled estimate was obtained using a random-effects model. Inspection of the individual reliability coefficients revealed substantial variation across the primary studies. The highest internal consistency estimate ($\alpha = .87$) was reported by González-Pizarro et al. (2023) in a large-scale study conducted with higher education students in Chile ($N = 549$). In contrast, the lowest reliability coefficient ($\alpha = .67$) was observed in the study by Fanchamps et al. (2021), which involved a relatively small sample of primary school students in the Netherlands ($N = 30$). The remaining reliability coefficients were distributed between these two extremes, as illustrated in the forest plot (Figure 3). Consistent with this observed variability, the heterogeneity analysis yielded a statistically significant Cochran's Q statistic, $Q_{(25)} = 150.69$, $p < .0001$, leading to the rejection of the homogeneity hypothesis. The I^2 statistic was estimated at 81.34%, indicating that a large proportion of the total variability in observed reliability coefficients is attributable to real between-study differences rather than sampling error. Supporting this interpretation, the estimated variance of the true effect sizes was $\tau^2 = 0.0409$ ($SE = 0.0173$), corresponding to a between-study standard deviation of $\tau = 0.2023$. Additionally, the H^2 value of 5.36 suggests that the observed variability in reliability estimates is more than five times greater than what would be expected under conditions of sampling variability alone (See Table 3).

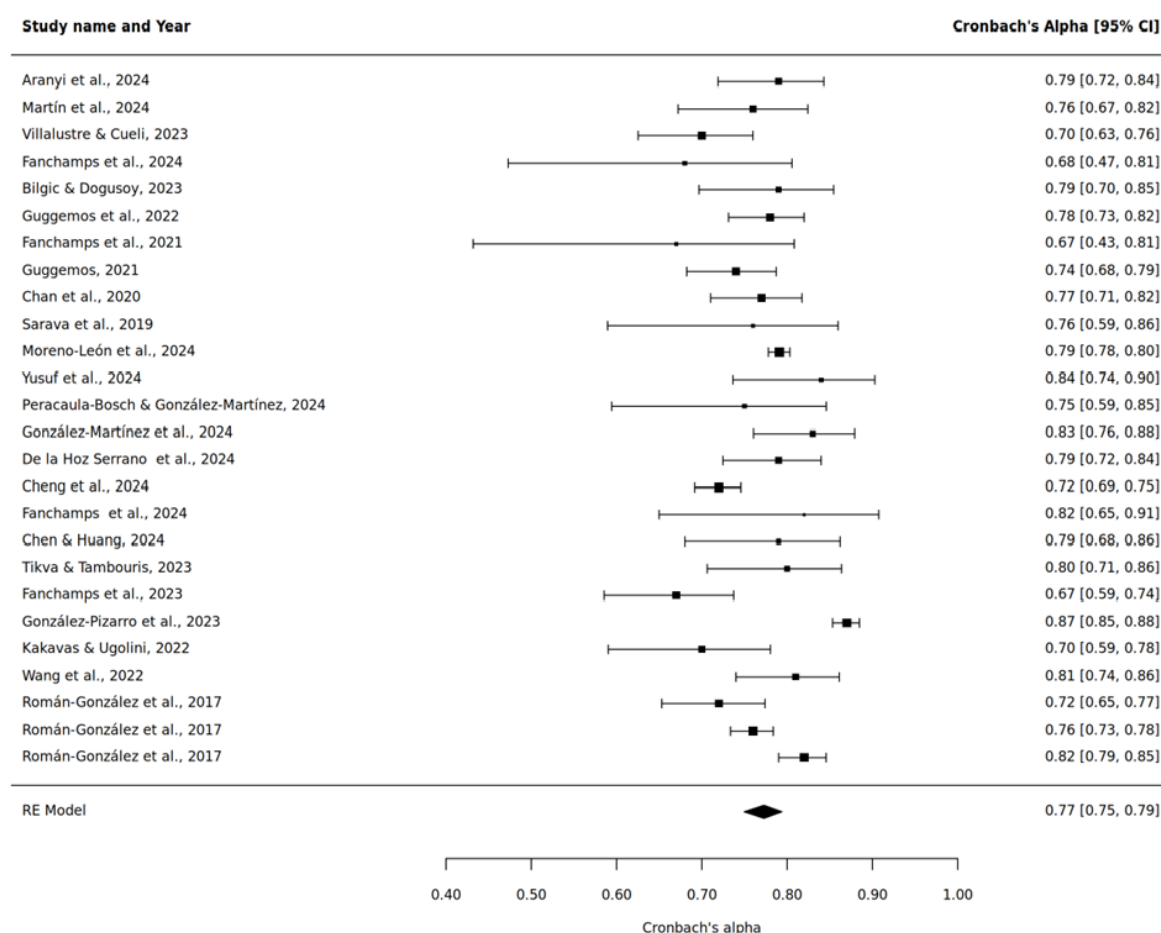


Figure 3. Forest plot

Table 3. Heterogeneity test results

τ^2 (SE)	τ	I^2 (%)	H^2	Q (df)	p-value	Pooled α	SE	z	95% CI
0.0409 (0.0173)	0.2023	81.34	5.36	150.69 (25)	< .0001	.773	0.011	68.71	[.750, .794]

Taken together, these indices provide convergent evidence of substantial between-study heterogeneity in the reported reliability coefficients. This pronounced heterogeneity empirically justifies the subsequent moderator analyses aimed at identifying potential sources of variation in the internal consistency estimates of the CTt across different study characteristics. Sensitivity analyses further indicated that the pooled reliability estimate was robust to the exclusion of small-sample studies. Excluding the single smallest study ($N = 21$) altered the pooled estimate negligibly ($\alpha = .772$, 95% CI [.748, .793]), as did excluding all studies with $N < 50$ ($\alpha = .774$, 95% CI [.748, .798]). A leave-one-out analysis confirmed that no single study shifted the pooled alpha by more than .008, indicating that the overall estimate was not disproportionately influenced by any individual sample.

Moderator Analysis

Given the substantial heterogeneity observed in the overall analysis, we conducted moderator analyses to identify potential sources of variability in the reliability estimates of the CTt. We examined both categorical moderators (using mixed-effects models analogous to ANOVA) and continuous moderators (using meta-regression). The results for the meta-regression analyses are presented in Table 3. Contrary to expectations, none of the examined continuous moderators exhibited a statistically significant relationship with the reliability coefficients. Specifically, sample size did not significantly predict reliability ($b = -0.00005$, $p = .6230$, $R^2 = 0.00\%$), suggesting that larger samples did not necessarily yield higher or more stable alpha coefficients in this dataset. Similarly, publication year ($p = .822$), percentage of female participants in the sample ($p = .315$), standard deviation of test scores ($p = .456$), and item number (treated as continuous; $p = .683$) failed to explain the heterogeneity in the reliability estimates. It should be noted that the moderator analyses for percentage of female participants ($k = 16$), standard deviation of test scores ($k = 14$), and urban/rural setting ($k = 16$) were based on considerably smaller

subsets of the full dataset due to missing data in the primary studies; consequently, these non-significant findings should be interpreted with caution, as the reduced sample sizes limit the statistical power to detect true moderating effects.

Table 4. Results of continuous moderators

Variable	k	b_j	p	R^2 (%)	QM	QE (df)	p (QE)
Sample Size	26	-0.00005	.6230	0.00	0.24	147.91 (24)	< .0001
Publication Year	26	-0.0087	.6727	0.00	0.18	149.83 (24)	< .0001
% Female Participants	16	-0.0748	.3151	1.05	1.01	93.33 (14)	< .0001
SD of Test Score	14	0.0040	.4561	0.00	0.56	34.52 (12)	.0006
Item Number	26	-0.0101	.5247	0.00	0.40	130.25 (24)	< .0001

Q_M : statistic to check if the model is well specified. b_j : regression coefficient of the moderating variable. R^2 : proportion of variance explained

The results for the categorical moderators are summarized in Table 5. Among the variables examined, school level showed the numerically largest association with reliability, although the omnibus test did not reach conventional statistical significance ($Q_M(3) = 6.98$, $p = .073$), accounting for 21.59% of the heterogeneity (R^2). Detailed analysis of the individual contrasts indicated that studies conducted with primary school students yielded numerically lower reliability estimates than the reference group of higher education, and this specific contrast remained statistically significant ($p = .0095$), even though the omnibus test across all school-level contrasts did not reach significance. Back-transformed to the original alpha metric, the estimated pooled reliability was $\alpha = .804$ (95% CI [.752, .845]) for higher education, $\alpha = .764$ (95% CI [.741, .785]) for lower secondary, $\alpha = .730$ (95% CI [.699, .759]) for primary, and $\alpha = .782$ (95% CI [.743, .814]) for upper secondary. No significant differences were found for lower secondary ($p = .359$) or upper secondary ($p = .388$) levels relative to higher education. In contrast, other categorical moderators did not significantly influence the reliability estimates. The language of administration ($Q_M(6) = 1.63$, $p = .950$), country ($Q_M(12) = 17.70$, $p = .125$), study setting (urban vs. rural; $Q_M(1) = 0.43$, $p = .510$), and item group (categorized as Original vs. Modified; $Q_M(1) = 1.79$, $p = .181$) were all non-significant. This implies that the CTt maintains relatively stable internal consistency across different languages, cultural contexts, and test versions within the scope of this meta-analysis. Tests of residual heterogeneity (QE) remained statistically significant for every moderator examined (all $p < .05$; see Tables 4 and 5), indicating that substantial unexplained variability persisted even after accounting for each moderator individually, and that unmeasured factors continue to contribute to the observed variability in alpha values.

Table 5. Results of categorical moderators

Moderator	k	df	QM	p	R^2 (%)	QE (df)	p (QE)	Result
School Level	25	3	6.98	.073	21.59	74.26 (21)	< .0001	ns
Language	17	6	1.63	.9502	45.26	30.88 (10)	.0006	ns
Country	26	12	17.70	.1250	37.79	39.15 (13)	.0002	ns
Urban/Rural	16	1	0.43	.5096	0.00	28.78 (14)	.0112	ns
Item Group (28 vs. Modified)	26	1	1.79	.1805	8.44	115.98 (24)	< .0001	ns

ns = non-significant

Because several categories within the Country and Language moderators were represented by only a single study, these variables were re-examined using more balanced groupings. When countries were aggregated into four broader regions (Europe, $k = 19$; Asia, $k = 5$; Americas, $k = 1$; Africa, $k = 1$), the omnibus test was nominally significant, $Q_M(3) = 15.07$, $p = .002$; however, this result was driven almost entirely by the two single-study categories (Africa and Americas), and should therefore be interpreted with caution. A more balanced binary contrast (Europe, $k = 19$, vs. Rest of World, $k = 7$) yielded only a marginal, non-significant trend, $Q_M(1) = 3.05$, $p = .081$. For Language, a binary contrast between Spanish ($k = 8$) and non-Spanish ($k = 9$) administrations—the only grouping with no single-study categories—showed no significant difference, $Q_M(1) = 0.01$, $p = .922$. Taken together, these supplementary analyses suggest that the evidence regarding country- and language-based variation in CTt reliability remains inconclusive given the current sparsity and geographic imbalance of the available literature, rather than providing clear support for either the presence or absence of cross-cultural variation.

Discussion

The CTt was developed by Román-González et al. (2017) to assess core dimensions of CT, drawing on the theoretical foundations proposed by Brennan and Resnick (2012) and Wing (2006). The instrument is designed to measure individuals' cognitive aptitude for formulating problems and solutions in a manner that can be executed

by a computer, with particular emphasis on fundamental computational concepts such as sequences, loops, and conditionals. Given the rapid dissemination and widespread adoption of the CTt across diverse educational contexts and countries, establishing the stability and generalizability of its psychometric properties has become increasingly important. Accordingly, the primary objective of the present reliability generalization meta-analysis was to estimate the overall reliability of CTt scores and to examine the extent to which reliability estimates vary across studies as a function of potential moderators.

The overall reliability generalization analysis yielded a pooled Cronbach's alpha coefficient consistent with the coefficients reported in the original validation study by Román-González et al. (2017) across different educational levels. This convergence between the meta-analytic estimate and the original validation results suggests that the CTt demonstrates stable and satisfactory internal consistency across diverse samples and over time. According to widely accepted psychometric standards, reliability coefficients above .70 are generally considered adequate for research purposes, whereas values exceeding .80 are preferred for applied or high-stakes assessments (Nunnally & Bernstein, 1994; George & Mallery, 2003). Within this framework, the present findings indicate that the CTt yields sufficiently reliable scores for assessing CT in educational research and practice. Despite this acceptable average level of internal consistency, the results revealed a substantial degree of heterogeneity among the included studies. This high level of between-study variability indicates that the reliability of the CTt is not consistent across different applications and sample characteristics. Such heterogeneity is consistent with previous reliability generalization research in educational psychology (Botella et al., 2010) and highlights the importance of reporting reliability estimates for each individual study, rather than relying exclusively on coefficients obtained in the original validation research (Vacha-Haase, 1998).

School level was the strongest of the moderators tested, though not in a way that survives the test as a whole: the omnibus comparison across the four educational levels falls just short of conventional significance. What does hold up is the contrast at the ends of the range: primary school students show reliably lower reliability than higher education students. Since these are also the two groups furthest apart developmentally, this is the comparison worth taking seriously rather than the omnibus test itself.

The direction is what a Piagetian account would predict: children in the concrete operational stage are still developing the visual-spatial and abstract reasoning that block-based items with nested loops or compound conditionals demand, and where that reasoning is still forming, guessing becomes more likely and measurement error rises accordingly (Sijtsma, 2009). Reading ability and digital fluency are also more variable among younger students, which adds construct-irrelevant variance on top of whatever the task itself demands (Messick, 1995). Each educational level is represented by only a handful of samples, which is not enough for the omnibus test to confirm that reliability declines in a graded way across all four levels rather than, say, dropping sharply at one particular transition. What the data support is narrower: reliability is not stable across age groups, and primary-level estimates in particular should be read with more caution than estimates from older samples. Settling whether the decline is gradual or concentrated at a specific developmental point will need reliability generalization work with denser sampling within each level.

In contrast, sample size did not emerge as a statistically significant predictor of reliability in the meta-regression analysis. This result contrasts with some earlier reliability generalization studies suggesting that larger sample sizes are associated with more stable reliability estimates (e.g., Charter, 1999). In the context of the CTt, the absence of a significant sample size effect may indicate that the internal structure of the test is sufficiently well-defined to yield comparable reliability estimates across studies, even when sample sizes are relatively small, provided that the measurement model is appropriate and the sample adequately represents the target population (Yurdugül, 2008). Furthermore, neither the language of test administration nor the country in which the study was conducted emerged as a significant moderator of reliability estimates in the primary analysis, although, as detailed in the Results, this conclusion warrants some caution given the sparsity of several category-level cells. With that caveat in mind, this pattern offers tentative support for the cross-cultural applicability of the CTt and is broadly consistent with the view that CT generalizes across cultural and linguistic contexts (Wing, 2006). The absence of strong language- and country-related effects is also consistent with, though not direct evidence for, the possibility that translation and adaptation processes into languages such as Spanish, German, and Turkish have preserved the psychometric properties of the original instrument, in line with established guidelines for test adaptation (Hambleton, 2005).

Interestingly, this pattern contrasts with Kiyici and Kahraman's (2022) reliability generalization of the self-report Computational Thinking Scale (Korkmaz et al., 2017), in which none of the examined moderators — including sample type — significantly predicted reliability. This divergence may reflect a meaningful distinction between performance-based CT assessments, such as the CTt, whose reliability appears sensitive to examinees'

developmental stage, and self-report CT scales, whose internal consistency may be comparatively more stable across sample characteristics. These findings also resonate with broader methodological patterns observed in CT-STEM intervention research. Some studies report contextual constraints such as substantial participant attrition and small subgroup sizes that limit statistical power (e.g., Leonard et al., 2016), while others are constrained to single-institution or single-country samples (Ayanwale & Omeh, 2026; Suherman et al., 2025; Tariq et al., 2025). Just as the reliability of CT measurement instruments such as the CTt varies systematically across educational and cultural contexts, the broader evidence base on CT development appears similarly shaped by contextual and sample-related constraints — reinforcing the need for both methodologically rigorous intervention designs and psychometrically sound assessment tools.

These findings carry particular relevance for STEM education, where CT has been formally incorporated as a core disciplinary practice. The Next Generation Science Standards designate computational thinking as one of the Science and Engineering Practices expected of K–12 students (NGSS Lead States, 2013), and instruments such as the CTt are increasingly used to evaluate whether STEM interventions succeed in building this competency, particularly within the computational problem-solving and data practices described by Weintrop et al. (2016). The reliability differences observed across school levels in the present study suggest that evaluators of STEM programs should not assume uniform measurement precision when using the CTt across the full K–12 range; conclusions drawn from primary-level data, in particular, may carry more measurement error than conclusions drawn from older cohorts. This has practical implications for STEM program evaluation and for teacher education efforts that rely on the CTt to document growth in computational thinking, since less precise measurement at younger ages could either mask real gains or produce misleading estimates of program impact.

Finally, the non-significant effect of the item group (original versus modified versions) indicates that adapted or shortened forms of the CTt tend to maintain acceptable levels of internal consistency. Nevertheless, it is noteworthy that the original 28-item version remains the most frequently employed form across the reviewed studies, which may partly account for the overall stability observed in reliability estimates.

Conclusion

Taken together, these findings reinforce the growing consensus in the psychometric literature that reliability is not an inherent property of a test, but rather is contingent upon the populations and contexts in which it is applied. Although the pattern observed for school level did not reach significance at the omnibus level, it still points to the importance of ensuring alignment between the cognitive demands of an assessment and the developmental characteristics of the examinees. Accordingly, caution is warranted when interpreting CTt scores obtained from primary education samples, and researchers and practitioners may benefit from complementing test scores with performance-based or alternative assessment approaches at this level. Overall, the present results highlight the necessity of reporting sample-specific reliability estimates and caution against the uncritical adoption of coefficients derived solely from original validation studies.

Scientific Ethics Declaration

* The authors declare that the scientific ethical and legal responsibility of this article published in JESEH journal belongs to the authors.

Conflict of Interest

* The authors declare that they have no conflicts of interest

Funding

*There is no funding received for this study.

References

- Ayanwale, M. A., & Omeh, C. B. (2026). AI-supported problem-based learning for enhancing computational thinking skills in STEM education. *Computers in Human Behavior: Artificial Humans*, 7, 100263.
- Barr, D., Harrison, J., & Conery, L. (2011). Computational thinking: A digital age skill for everyone. *Learning & Leading with Technology*, 38(6), 20-23.
- Bati, K. (2022). A systematic literature review regarding computational thinking and programming in early childhood education. *Education and Information Technologies*, 27(2), 2059-2082.
- Begg, C. B., & Mazumdar, M. (1994). Operating characteristics of a rank correlation test for publication bias. *Biometrics*, 50(4), 1088-1101.
- Belletini, C., Lonati, V., Malchiodi, D., Monga, M., Morpurgo, A., and Torelli, M. (2015). "How challenging are Bebras' tasks? An IRT analysis based on the performance of Italian students," in Proceedings of the 2015 ACM Conference on Innovation and Technology in Computer Science Education, ITiCSE '15 (New York, NY: Association for Computing Machinery), 27-32. doi: 10.1145/2729094.2742603
- Beretvas, S. N., & Pastor, D. A. (2003). Using mixed-effects models in reliability generalization studies. *Educational and Psychological Measurement*, 63(1), 75-95.
- Bocconi, S., Chiocciariello, A., Dettori, G., Ferrari, A., Engelhardt, K., Kampylis, P., & Punie, Y. (2016, June). Developing computational thinking: Approaches and orientations in K-12 education. In *EdMedia+ innovate learning* (pp. 13-18). Association for the Advancement of Computing in Education (AACE).
- Bonett, D. G. (2002). Sample size requirements for testing and estimating coefficient alpha. *Journal of Educational and Behavioral Statistics*, 27(4), 335-340. <https://doi.org/10.3102/10769986027004335>
- Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2010). A basic introduction to fixed-effect and random-effects models for meta-analysis. *Research Synthesis Methods*, 1(2), 97-111.
- Botella, J., Suero, M., & Gambara, H. (2010). Psychometric inferences from a meta-analysis of reliability and internal consistency coefficients. *Psychological Methods*, 15(4), 386-397.
- Brennan, K., & Resnick, M. (2012). *New frameworks for studying and assessing the development of computational thinking*. In *Proceedings of the American Educational Research Association Annual Meeting* (pp. 1-25). Vancouver, Canada.
- Charter, R. A. (1999). Sample size requirements for precise estimates of reliability coefficients. *Journal of Clinical and Experimental Neuropsychology*, 21(4), 559-566.
- Chen, G., Shen, J., Barth-Cohen, L., Jiang, S., Huang, X., and Eltoukhy, M. (2017). Assessing elementary students' computational thinking in everyday reasoning and robotics programming. *Comput. Educ.* 109, 162-175. doi: 10.1016/j.compedu.2017.03.001
- Duval, S., & Tweedie, R. (2000). Trim and fill: A simple funnel-plot-based method of testing and adjusting for publication bias in meta-analysis. *Biometrics*, 56(2), 455-463.
- Egger, M., Davey Smith, G., Schneider, M., & Minder, C. (1997). Bias in meta-analysis detected by a simple, graphical test. *BMJ*, 315(7109), 629-634.
- El-Hamamsy, L., Zapata-Cáceres, M., Barroso, E. M., Mondada, F., Zufferey, J. D., et al. (2022). The competent computational thinking test: Development and validation of an unplugged computational thinking test for upper primary school. *J. Educ. Comput. Res.* 60:07356331221081753.
- Falloon, G. (2024). Advancing young students' computational thinking: An investigation of structured curriculum in early years primary schooling. *Computers & Education*, 216, 105045.
- Gane, B. D., Israel, M., Elagha, N., Yan, W., Luo, F., and Pellegrino, J. W. (2021). Design and validation of learning trajectory-based assessments for computational thinking in upper elementary grades. *Computer Science Education* 31, 141-168.
- George, D., & Mallery, P. (2003). *SPSS for Windows step by step: A simple guide and reference* (4th ed.). Allyn & Bacon.
- Grover, S., & Pea, R. (2013). Computational thinking in K-12: A review of the state of the field. *Educational Researcher*, 42(1), 38-43.
- Hambleton, R. K. (2005). Issues, designs, and technical guidelines for adapting tests into multiple languages and cultures. In R. K. Hambleton, P. F. Merenda, & C. D. Spielberger (Eds.), *Adapting educational and psychological tests for cross-cultural assessment* (pp. 3-38). Lawrence Erlbaum Associates.
- Higgins, J. P. T., Thompson, S. G., Deeks, J. J., & Altman, D. G. (2003). Measuring inconsistency in meta-analyses. *BMJ*, 327(7414), 557-560.
- Hubwieser, P., and Mühling, A. (2014). "Playing PISA with bebras," in *Proceedings of the 9th Workshop in Primary and Secondary Computing Education, WiPSCE '14* (New York, NY: Association for Computing Machinery), 128-129.
- Kassambara, A. (2023). *rstatix: Pipe-friendly framework for basic statistical tests* (Version 0.7.2) [R package]. Comprehensive R Archive Network (CRAN). <https://CRAN.R-project.org/package=rstatix>
- Kiyici, G., & Kahraman, N. (2022). A meta-analytic reliability generalization study of the computational thinking scale. *Science Insights Education Frontiers*, 13(2), 1859-1874. <https://doi.org/10.15354/sief.22.ma011>

- Kong, S.-C., and Lai, M. (2022). Validating a computational thinking concepts test for primary education using item response theory: an analysis of students' responses. *Comput. Educ.* 187:104562.
- Kuder, G. F., & Richardson, M. W. (1937). The theoretical estimation of test reliability. *Psychometrika*, 2(3), 151–160.
- Lee, J., Joswick, C., & Pole, K. (2023). Classroom play and activities to support computational thinking development in early childhood. *Early Childhood Education Journal*, 51(3), 457–468.
- Leonard, J., Buss, A., Gamboa, R., Mitchell, M., Fashola, O. S., Hubert, T., & Almughyirah, S. (2016). Using robotics and game design to enhance children's self-efficacy, STEM attitudes, and computational thinking skills. *Journal of Science Education and Technology*, 25(6), 860–876.
- Li, Y., Xu, S., and Liu, J. (2021). Development and validation of computational thinking assessment of Chinese elementary school students. *J. Pacific Rim Psychol.* 15:183449092110102.
- Marín-Martínez, F., & Sánchez-Meca, J. (2010). Weighting by inverse variance or by sample size in random-effects meta-analysis. *Educational and Psychological Measurement*, 70(1), 56–73.
- Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist*, 50(9), 741–749.
- Müller, K., & Wickham, H. (2023). *tibble: Simple data frames* (Version 3.2.1) [R package]. Comprehensive R Archive Network (CRAN). <https://CRAN.R-project.org/package=tibble>
- NGSS Lead States. (2013). *Next Generation Science Standards: For states, by states*. The National Academies Press.
- Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory* (3rd ed.). McGraw-Hill.
- Ooms, J. (2024). *writexl: Export data frames to Excel xlsx format* (Version 1.5.0) [R package]. Comprehensive R Archive Network (CRAN). <https://CRAN.R-project.org/package=writexl>
- Papert, S. (1980). *Mindstorms: Children, computers, and powerful ideas*. New York: Basic Books.
- Parker, M., Kao, Y., Saito-Stehberger, D., Franklin, D., Krause, S., Richardson, D., et al. (2021). “Development and preliminary validation of the assessment of computing for elementary students (ACES),” in SIGCSE '21: Proceedings of the 52nd ACM Technical Symposium on Computer Science Education (Virtual Event), 10–16.
- R Core Team. (2024). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. <https://www.R-project.org/>
- Relkin, E., and Bers, M. (2021). “Techcheck-k: a measure of computational thinking for kindergarden children,” in 2021 *IEEE Global Engineering Education Conference (EDUCON)* (Vienna), 1696–1702.
- Relkin, E., de Ruiter, L., and Bers, M. U. (2020). TechCheck: development and validation of an unplugged assessment of computational thinking in early childhood education. *J. Sci. Educ. Technol.* 29, 482–498.
- 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.
- Rosenthal, R. (1979). The file drawer problem and tolerance for null results. *Psychological Bulletin*, 86(3), 638–641.
- Shute, V. J., Sun, C., & Asbell-Clarke, J. (2017). Demystifying computational thinking. *Educational Research Review*, 22, 142–158.
- Suherman, S., Vidákovich, T., Mujib, M., Hidayatulloh, H., Andari, T., & Susanti, V. D. (2025). The role of STEM teaching in education: An empirical study to enhance creativity and computational thinking. *Journal of Intelligence*, 13(7), 88.
- Sijtsma, K. (2009). On the use, the misuse, and the very limited usefulness of Cronbach's alpha. *Psychometrika*, 74(1), 107–120.
- Tariq, R., Aponte Babines, B. M., Ramirez, J., Alvarez-Icaza, I., & Naseer, F. (2025). Computational thinking in STEM education: Current state-of-the-art and future research directions. *Frontiers in Computer Science*, 6, 1480404.
- Thompson, B., & Vacha-Haase, T. (2000). Psychometrics is datametrics: The test is not reliable. *Educational and Psychological Measurement*, 60(2), 174–195.
- Tsarava, K., Moeller, K., Román-González, M., Golle, J., Leifheit, L., Butz, M. V., & Ninaus, M. (2022). A cognitive definition of computational thinking in primary education. *Computers & Education*, 179, 104425.
- Vacha-Haase, T. (1998). Reliability generalization: Exploring variance in measurement error affecting score reliability across studies. *Educational and Psychological Measurement*, 58(1), 6–20.
- Vee, A. (2017). *Coding literacy: How computer programming is changing writing*. Mit Press.
- Veroniki, A. A., Jackson, D., Viechtbauer, W., Bender, R., Bowden, J., Knapp, G., Kuss, O., Higgins, J. P. T., Langan, D., & Salanti, G. (2016). Methods to estimate the between-study variance and its uncertainty in meta-analysis. *Research Synthesis Methods*, 7(1), 55–79.

- Viechtbauer, W. (2010). Conducting meta-analyses in R with the *metafor* package. *Journal of Statistical Software*, 36(3), 1–48.
- Weintrop, D., Beheshti, E., Horn, M., Orton, K., Jona, K., Trouille, L., & Wilensky, U. (2016). Defining computational thinking for mathematics and science classrooms. *Journal of Science Education and Technology*, 25(1), 127–147.
- Wickham, H. (2016). *ggplot2: Elegant graphics for data analysis*. Springer.
- Wickham, H., & Bryan, J. (2023). *readxl: Read Excel files* (Version 1.4.3) [R package]. Comprehensive R Archive Network (CRAN). <https://CRAN.R-project.org/package=readxl>
- Wickham, H., François, R., Henry, L., & Müller, K. (2022). *dplyr: A grammar of data manipulation* (Version 1.0.10) [R package]. Comprehensive R Archive Network (CRAN). <https://CRAN.R-project.org/package=dplyr>
- Wickham, H., Vaughan, D., & Girlich, M. (2024). *tidyr: Tidy messy data* (Version 1.3.1) [R package]. Comprehensive R Archive Network (CRAN). <https://CRAN.R-project.org/package=tidyr>
- Wing, J. M. (2006). Computational thinking. *Communications of the ACM*, 49(3), 33–35.
- Yang, W., Ng, D. T. K., & Su, J. (2023). The impact of story-inspired programming on preschool children's computational thinking: A multi-group experiment. *Thinking Skills and Creativity*, 47, 101218.
- Yurdugül, H. (2008). Minimum sample size for Cronbach's coefficient alpha: A Monte-Carlo study. *Hacettepe University Journal of Education*, 35, 397–405.
- Zapata-Cáceres, M., Martín-Barroso, E., and Román-González, M. (2020). “Computational thinking test for beginners: design and content validation,” in *2020 IEEE Global Engineering Education Conference (EDUCON)* (Porto), 1905–1914.

Author(s) Information

Seyma Irmak

Amasya University, Faculty of Education, Department of
Mathematics and Science Education, Amasya, Türkiye
ORCID: <https://orcid.org/0000-0003-3831-8244>

Kaan Bati

Hacettepe University, Faculty of Education, Department of
Mathematics and Science Education, Ankara, Türkiye
Contact e-mail: kaanbati@gmail.com
ORCID: <https://orcid.org/0000-0002-6169-7871>

Appendix: Included Studies in this Review

- *Aranyi, G., Kovács, K., Kemény, F., Pachner, O., Klein, B., & Remete, E. P. (2024). Computational thinking in university students: The role of fluid intelligence and visuospatial ability. *PLOS ONE*, 19(8), e0309412.
- *Chan, S.-W., Looi, C.-K., & Sumintono, B. (2021). Assessing computational thinking abilities among Singapore secondary students: A Rasch model measurement analysis. *Journal of Computers in Education*, 8(2), 213–236.
- *Chen, C.-M., & Huang, M.-Y. (2024). Enhancing programming learning performance through a jigsaw collaborative learning method in a metaverse virtual space. *International Journal of STEM Education*, 11, 36.
- *Cheng, M., Lai, X., Tao, D., et al. (2024). Children's programming environment acceptance: Extending the boundary conditions to programming competition, computational thinking, and programming modality. *Education and Information Technologies*, 29, 939–969.
- *De la Hoz Serrano, A., Melo Niño, L. V., Álvarez Murillo, A., Martín Tardío, M. Á., Cañada Cañada, F., & Cubero Juárez, J. (2024). Analysis of gender issues in computational thinking approach in science and mathematics learning in higher education. *European Journal of Investigation in Health, Psychology and Education*, 14(11), 2865–2882.
- *Fanchamps, N., Van Gool, E., Slangen, L., & Hennissen, P. (2024). The effect on computational thinking and identified learning aspects: Comparing unplugged smartGames with SRA-Programming with tangible or on-screen output. *Education and Information Technologies*, 29(3), 2999–3024.
- *Fanchamps, N., Van Gool, E., Folkertsma, A., & De Meyst, K. (2024). The influence of music producing and creativity on computational thinking in primary school children. *Education Sciences*, 14(12), 1380.
- *González-Pizarro, F., López, C., Vásquez, A., & Castro, C. (2023). Inequalities in computational thinking among incoming students in a STEM Chilean university. *IEEE Transactions on Education*, 67(2), 180–189.
- *González-Martínez, J., Peracaula-Bosch, M., & Meyerhofer-Parra, R. (2024). Impact of intensive programming training on the development of computational thinking in prospective teachers [Impacto de una formación intensiva en programación en el desarrollo del pensamiento computacional en futuros/as maestros/as]. *RIED—Revista Iberoamericana de Educación a Distancia*, 27(1).
- *Guggemos, J. (2021). On the predictors of computational thinking and its growth at the high-school level. *Computers & Education*, 161, 104060.
- *Kakavas, P., & Ugolini, F. C. (2022). *Psychometric testing of the adapted Computational Thinking Test in Greek culture: Ensuring its reliability, structure and convergent validity*. In ICERI 2022 Proceedings (pp. 4281–4289). International Conference of Education, Research and Innovation.
- *Martín, L. C. S., Hijón-Neira, R., Pizarro, C., & Cañas, J. M. (2024). Fostering computational thinking with simulated 3D robots in secondary education. *Computer Applications in Engineering Education*, 32(4), e22740.
- *Moreno-León, J., Vasco-González, M., Román-González, M., & Robles, G. (2024). Investigating the impact of programming activities on computational thinking and AI literacy in Spanish schools. In *Proceedings of the 19th Conference on Primary and Secondary Computing Education Research (WiPSCE '24)*, Article 7, pp. 1–10. Association for Computing Machinery.
- *Peracaula-Bosch, M., & González-Martínez, J. (2024). Developing computational thinking among preservice teachers. In *ATEE Spring Conference book of abstracts* (pp. 57–61).
- *Tikva, C., & Tambouris, E. (2023). The effect of scaffolding programming games and attitudes towards programming on the development of computational thinking. *Education and Information Technologies*, 28, 6845–6867.
- *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.
- *Tsarava, K., Leifheit, L., Ninaus, M., Román-González, M., Butz, M. V., Golle, J., Trautwein, U., & Moeller, K. (2019). Cognitive correlates of computational thinking: Evaluation of a blended unplugged/plugged-in course. In *Proceedings of the 14th Workshop in Primary and Secondary Computing Education (WiPSCE '19)*, Article 24, pp. 1–9. Association for Computing Machinery.
- *Wang, D., Luo, L., Luo, J., Lin, S., & Ren, G. (2022). Developing computational thinking: Design-based learning and interdisciplinary activity design. *Applied Sciences*, 12(21), 11033.
- *Villalustre, L., & Cueli, M. (2023). Assessing the computational thinking of pre-service teachers: A gender and robotics programming experience analysis. *Education Sciences*, 13(10), 1032.
- *Yusuf, A., Noor, N. M., & Román-González, M. (2024). Interaction patterns during block-based programming activities predict computational thinking: Analysis of the differences in gender, cognitive load, spatial ability, and programming proficiency. In *AI, computer science and robotics technology*. IntechOpen.