Research Article https://doi.org/10.12973/ijem.8.4.687

International Journal of Educational Methodology
Volume 8, Issue 4, 687 - 698.
ISSN: 2469-9632
http://www.ijem.com/

# Rethinking the Components of Regulation of Cognition through the Structural Validity of the Meta-Text Test

**Marcio Alexander Castillo-Diaz*** 
Universidad Nacional Autónoma de Honduras, HONDURAS

**Cristiano Mauro Assis Gomes**
Universidade Federal de Minas Gerais, BRAZIL

**Enio Galinkin Jelihovschi**
Universidade Estadual de Santa Cruz, BRAZIL

*Received: June 14, 2022 ▪ Revised: September 2, 2022 ▪ Accepted: October 19, 2022*

**Abstract:** The field of studies in metacognition points to some limitations in the way the construct has traditionally been measured and shows a near absence of performance-based tests. The Meta-Text is a performance-based test recently created to assess components of cognition regulation: planning, monitoring, and judgment. This study presents the first evidence on the structural validity of the Meta-Text, by analyzing its dimensionality and reliability in a sample of 655 Honduran university students. Different models were tested, via item confirmatory factor analysis. The results indicated that the specific factors of planning and monitoring do not hold empirically. The bifactor model containing the general cognition regulation factor and the judgment-specific factor was evaluated as the best model (CFI = .992; NFI = .963; TLI = .991; RMSEA = .021). The reliability of the factors in this model proved to be acceptable (Ω = .701 & .699). The judgment items were well loaded only by the judgment factor, suggesting that the judgment construct may actually be another component of the metacognitive knowledge dimension but having little role in cognition regulation. The results show initial evidence on the structural validity of the Meta-Text and give rise to information previously unidentified by the field which has conceptual implications for theorizing metacognitive components.

**Keywords:** *Metacognition, performance-based testing, regulation of cognition, structural validity.*

**To cite this article:** Castillo-Diaz, M. A., Gomes, C. M. A., & Jelihovschi, E. G. (2022). Rethinking the components of regulation of cognition through the structural validity of the meta-text test. *International Journal of Educational Methodology*, *8*(4), 687-698. https://doi.org/10.12973/ijem.8.4.687

## Introduction

Metacognition is an ability that plays a key role in general high-order cognitive functioning, it is associated to processes of learning self-regulation (e.g., Panadero, 2017; Schunk & Greene, 2018), executive functions (e.g., Filippi et al., 2020; Roebers, 2017), critical and complex thinking (e.g., Amin et al., 2020; Silva & Iturra, 2021), creativity (e.g., Jia et al., 2019; Preiss et al., 2019), among others. In educational settings, metacognition is also associated with educational outcomes, such as performance, in such a way that its components are targeted for diagnosis, training, and intervention (Cromley & Kunze, 2020; Saenz et al., 2019).

The literature points out that metacognition is primarily composed of two major components or domains: metacognitive knowledge and cognition regulation (Craig et al., 2020). Both domains interact with each other. The former refers to what people know about their own functioning, what implies a knowledge about how they could engage more efficiently with a specific task. In turn, cognition regulation refers to the ability to control and monitor the cognitive strategies used in task performance (Azevedo, 2020; Muijs & Bokhove, 2020; Norman et al., 2019).

There is an almost complete predominance of measuring metacognitive components by self-report instruments and think aloud protocols (Craig et al., 2020; Gascoine et al., 2017; Ohtani & Hisasaka, 2018). However, the literature points out important limitations of these measures, especially with regard to the measurement of the cognition regulation domain. Cognition regulation involves the "online" or "in-moment" process as the cognitive task is performed, nevertheless, the self-report instruments make it difficult to accurately assess the domain, since these instruments are usually applied offline, i.e., before or after task performance (Akturk & Sahin, 2011; Craig et al., 2020). Moreover, there

* **Corresponding author:**
Marcio Alexander Castillo-Diaz, Universidad Nacional Autónoma de Honduras, Student Affairs Department/ Graduate Program in Psychometrics and Educational Evaluation, Honduras. ✉ marcio.castillo@unah.edu.hn

are several evidences about the biases involved in self-report instruments, such as social desirability, acquiescence, and the possible lack of knowledge of respondents about their own cognitive processes (Abernethy, 2015; Wetzel et al., 2016). In turn, think aloud protocols allow an online and more accurate gauging of the processes of cognition regulation involved in the execution of cognitive tasks (Hu & Gao, 2017). However, this type of measurement requires the presence of judges to evaluate the protocols, bringing significant risks of confirmatory bias in the measurement process (Greene et al., 2018; Wolcott & Lobczowski, 2021). Furthermore, the use of think aloud protocols requires an intensive individual assessment process, resulting in costly studies with small samples (e.g., Van der Stel & Veenman, 2008; Veenman & Van Cleef, 2018).

Another way to address the problem of measuring metacognition is the construction and validation of performance-based tests. This type of test allows an assessment of the construct at the moment the task is performed by the respondent, allowing scores to be obtained via their performance (Castillo-Diaz & Gomes, 2021; Gomes, Araujo & Castillo-Diaz, 2021). Furthermore, the use of performance tests does not require the participation of judges to evaluate the scores. This substantially reduces the confirmatory bias present in think aloud and allows for the generation of less costly studies that can be done on large samples.

Ohtani and Hisasaka's (2018) meta-analysis shows that performance-based measures are far superior to self-report measures. The correlations of the self-report metacognitive measures with academic performance (r = .18; 95% confidence interval = .13-.22) are much lower than the correlations of this outcome with the performance-based measures of the think aloud method tasks (r = .41; 95% confidence interval = .31-.52). However, the think aloud method is not supported by tests, and is therefore often applied only to small samples, relying heavily on judges for the production of scores and validation of constructs. Therefore, the development of performance-based tests would allow investigating in much larger samples the real predictive power of metacognitive components in relation to educational outcomes (e.g., Castillo-Diaz & Gomes, 2022; Gomes et al., 2014). The use of better measures, without significant noise or biases, has important practical applications, since better evidence, arising from these measures, permits to design more pertinent diagnoses and educational intervention strategies (Donker et al., 2014; Jansen et al., 2019).

The dominance in the design and use of self-report instruments and the absence of performance-based testing is not a prerogative of the field of metacognition. The field of studies in learning approaches suffers from this same handicap. Nevertheless, we observe an initial effort in the development of performance-based tests in this field, envisioning new possibilities in measuring and building evidence for the area (e.g., Gomes et al., 2020, 2021; Gomes & Nascimento, 2021). In the case of metacognition, to the best of our knowledge, few tests are identified that aim to measure the construct through respondent performance (Desoete et al., 2001; Golino & Gomes, 2011; Neuenhaus et al., 2011). When analyzing which of these tests present validity studies based on their factor structure, we found that only the Metacognitive Monitoring Test (MMT, also called Reading Monitoring Test or Read Monitoring Test) presents such evidence (Castillo-Diaz & Gomes, 2022; Golino & Gomes, 2011; Gomes & Golino, 2014; Gomes et al., 2014). Moreover, the MMT is the only performance-based test that was developed to be applied to higher education students. However, the MMT does show some limitations. These are: (1) the measurement of a single metacognitive component, (2) the borderline reliability (between values of .60 and .70 in the alpha and omega indices), (3) and above all the intensive work for generating the item scores, since the respondents' justification must be read and evaluated to produce the scores.

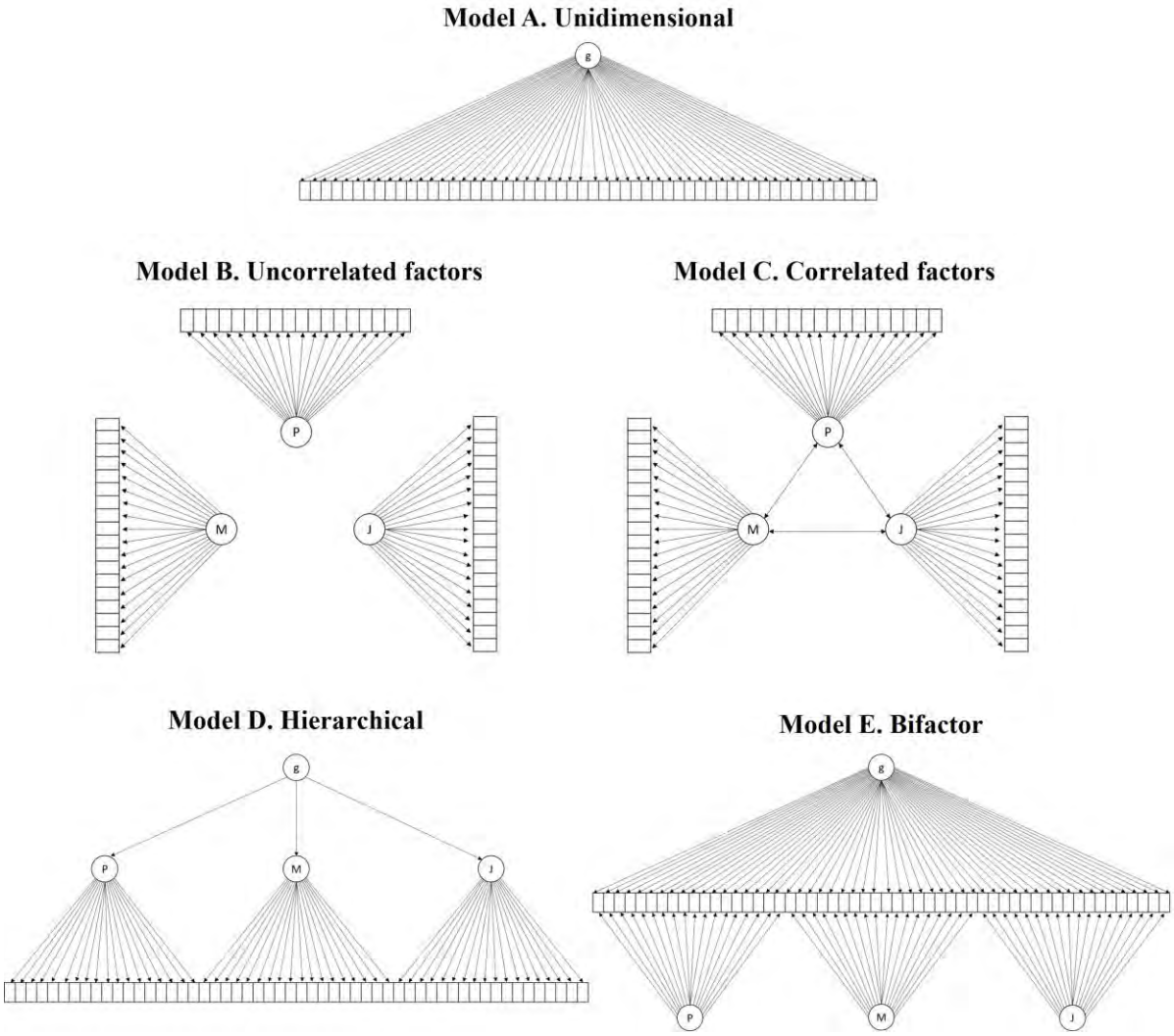
The Meta-Performance battery is a performance-based assessment instrument recently created in order to assess three specific metacognitive abilities from the regulation domain of cognition: planning, monitoring, and judgment (Castillo-Diaz & Gomes, 2021). Planning is conceived in the battery as a very specific metacognitive ability, involving only the individual's ability to properly identify specific sequences of steps that allow the resolution of a task (Oliveira & Nascimento, 2014). Monitoring is also defined as a very specific metacognitive ability, that is, the individual's ability to detect errors at the moment they are performing a given task (e.g., Golino & Gomes, 2011; Pires & Gomes, 2018). Judgment is defined by the test authors within the concurrent judgment paradigm, it is the person's evaluation of their own performance after solving each test item (Schraw, 2009). The metacognitive abilities involved in the battery are relevant predictors to education, insofar as meta-analyses indicate that, compared to the metacognitive knowledge, cognition regulation is the domain of greatest importance in terms of predicting student academic performance (Dent & Koenka, 2016; Ohtani & Hisasaka, 2018).

The battery is composed of two tests. The first assesses metacognition in reading comprehension tasks (Meta-Text), while the second makes this assessment in tasks of solving arithmetic expressions (Meta-Number). Previous research indicates favorable evidence for the content validity of the battery (Castillo-Diaz & Gomes, 2021). However, the tests of the battery have not yet undergone analysis of their structural validity.

*Tested Models*

In this paper, five models are tested on the dimensionality of the Meta-Text test (Figure 1). Model A (Unidimensional) establishes that only the cognition regulation domain explains the variance of all test items, rejecting the presence of any of the specific metacognitive abilities of planning, judgment, and monitoring. Model B (uncorrelated factors)

establishes the presence of the three specific metacognitive abilities, and assumes that these abilities are independent, it does not accept the possibility of the presence of the cognition regulation domain, insofar as metacognitive theory assumes that these abilities are related because they are part of this domain. Model C (correlated factors) assumes that the three specific metacognitive abilities correlate. However, this model does not explicitly establish the presence of the cognition regulation domain as an explanation of the correlations between the metacognitive abilities. Model D (Hierarchical), also called second-order, defines that the cognition regulation domain explains the correlations between specific metacognitive abilities. Finally, Model E (Bifactor) assumes that both the cognition regulation domain and the specific metacognitive abilities directly explain people's performance on the items.




*Note: g = Regulation of cognition; P = Planning; M = Monitoring; J = Judgment.*

*Figure 1. Analyzed Models of the Dimensionality of the Meta-Text Test*

One of the advantages of performance-based tests is that they allow for optimal testing of the empirical plausibility of constructs. Self-report tests also allow this to be tested, but performance-based items allow constructs to be empirically tested through performance, which makes the evidence more robust than evidence supported by respondents' self-report via reading the wording of items that represent certain behaviors. The very construction of self-report items permits the respondent to perceive an association between certain items and respond accordingly. Performance-based items, on the other hand, do not depend on this perception of the respondent, so this type of bias is not relevant for this type of items. The think aloud method, although based on performance, is based on judges' scores, so that, the testing construct is quite fragile, despite having some validity. This type of analysis has a strong potential for bias, because judges, despite all methodological precautions to avoid positive bias, generate scores that potentially produce associated scores among task components based on the judges' prior expectations. In our paper, besides testing the structural validity of the Meta-Text test, we are also testing the validity of the metacognitive abilities of planning, monitoring, judgment, and regulation of cognition. The different models analyzed in this study allow us to refute or corroborate these components of metacognition. As highlighted, the almost exclusive dominance of self-report tests and think aloud procedures has brought weak evidence on the empirical plausibility of these constructs and our study

permits the generation of evidence based on a more solid methodology which are not employed by almost all studies in the field.

## Methodology

*Sample*

The participants of this study were selected through a convenience sampling that included 655 higher education students from the largest public university in Honduras in Central America. The sample is characterized by a predominance of females (N = 441; 67.3%), young adults (M = 20.14 years; SD = 3.06), and belonging to the central campus (N = 489; 74.7%). The sample has 274 (41.8%) students from the economic sciences, 159 (24.3%) students from the social sciences, humanities, and arts, 110 (16.8%) students from the exact sciences and engineering, and 112 (17.1%) from the biological and health sciences.

*Instruments*

Meta-Text Test. This is a performance-based test that is part of the Meta-Performance Battery, recently created with the goal of assessing specific abilities in the domain of cognition regulation. Details of the battery and items are found in Castillo-Diaz and Gomes (2021). The Meta-Performance battery is a production of the Laboratory for Research on Cognitive Architecture (Laboratório de Investigação da Arquitetura Cognitiva [LAICO]), which has a mission to articulate psychometrics to educational psychology and to build a broad set of tests on predictors of educational outcomes. The Monitoring Metacognitive Test (MMT) itself, cited in this article, was developed by LAICO and also shows good evidence of internal and external validity from several studies (e.g., Golino & Gomes, 2011; Gomes & Golino, 2014; Gomes et al., 2014). The MMT is a well-established metacognitive test in Brazil that has influenced the creation of the Meta-Performance Battery (Castillo-Diaz & Gomes, 2021) and a methodology for measuring metacognition in school and academic assessments (Gomes, 2021; Pires & Gomes, 2017, 2018).

The Meta-Text is structured by a set of 18 questions that were carefully designed with the purpose of covering a thematic diversity of texts and not requiring relevant prior knowledge of the respondent about these topics. Each question is composed of three elements: (a) a statement describing a hypothetical author's goal for writing a text; (b) five possible sentences for composing the text, depending on the presented goal, and; (c) a text written by the hypothetical author, using some of the available sentences. The respondent's task is to answer three specific commands: Command A measures planning ability and therefore asks the respondent to create a plan regarding what would be the appropriate sentences for the author to adequately achieve his objective. Command B measures the judgment ability and therefore asks the respondent to evaluate his/her own planning, indicating whether he/she thinks he/she answered Command A correctly or incorrectly. Command C measures the monitoring ability and asks the respondent to analyze the text written by the hypothetical author, identifying if there are errors in the text, either by the presence of sentences that do not contribute to the author's objective or omitted sentences that should have been written (Castillo-Diaz & Gomes, 2021).

Each test command represents one item, making up 18 items for each of the abilities measured: planning (Command A), judgment (Command B), and monitoring (Command C), for a total of 54 items. Each item of Command A and C has a score of one point (1) if the answer is correct and zero points (0) if it is wrong. In the case of the judgement (Command B), if the respondent thinks he has got command A right of a certain question, then their score will be one (1) for the judgement; otherwise their score will be zero (0). This judgment score does not indicate the accuracy of the respondents' judgment. We created a calculation for the accuracy that is presented in the data analysis section. The test is designed to be completed in 60 minutes, which is the maximum duration.

*Procedures*

Undergraduate students, with active enrollment from different areas of knowledge at the National Autonomous University of Honduras (UNAH) were contacted by e-mail and invited to participate in the research. The invitation was sent via e-mails registered in the university's registration system database. Data collection was conducted online via Google Forms during the first quarter of the year 2021. Students were invited to participate in the study on a voluntary basis, receiving prior information about the purpose, procedures, and their rights to withdraw from participating in the research at any time. Their participation was conditioned to the acceptance of a Free and Informed Consent Form (FICF). Only data from students who previously agreed to the FICF were considered. Students filled out a form that included information about sociodemographic data and the Meta-Text Test. According to the analyses obtained in a pilot application of the test, students were expected to take between 40 and 60 minutes to complete it. However, there was no time restriction for taking the test on the online platform. Approval for this study was obtained from the Institutional Review Board of the University's Dean of Student Affairs.

*Data Analysis*

The first stage of the analyses involved the description of descriptive statistics about the difficulty of the Meta-Text Test items. In the case of the judgment items, accuracy was calculated as follows: if the person got the planning item right and judged that they got this item right or if they got the planning item wrong and judged that they got this item wrong, then they got the judgment item right, in terms of accuracy, because they judged their performance correctly. In turn, if the person missed the planning item and thinks they got this item right, or if they get the planning item right and think they got this item wrong, then they get the judgment item wrong, in terms of accuracy. When a person gets the item right, their accuracy score is 1; otherwise, their accuracy score is 0.

The second step involved structural validity analysis. Different models, representing different factor structures, were tested using confirmatory factor analysis of items. All models analyzed included the presence of covariance between pairs of planning and monitoring items linked to the same question. Models with the following specifications were analyzed: Model A (Unidimensional) states that the latent variable "regulation of cognition" explains the variance of the 54 test items. Model B (uncorrelated factors) states that three latent variables "planning", "monitoring" and "judgment" explain a set of 18 items each. In this model the latent variables are orthogonal, that is, the correlation between them is fixed at zero. Model C (correlated factors) is just a variant of Model B, since it has the same latent variables, but allows them to correlate. Model D (Hierarchical) is composed of the same latent variables as models B and C, but incorporates a second-level general factor that explains the covariance between the first-level latent variables. Model E (Bifactor) has the same variables as Model D, but defines that the general factor of cognition regulation directly explains the 54 test items. In the bifactor model all latent variables are orthogonal, constraining their correlation to zero (Reise, 2012). Figure 1 presents the structure of the tested models. For parsimony purposes, covariance is not drawn between pairs of planning and monitoring items linked to the same question.

Since the test item scores have a dichotomous nature (correct and incorrect), the Weighted Least Square Mean and Variance Adjusted (WLSMV) estimator was used. According to the literature on confirmatory factor analysis, the WLSMV is the best alternative for modelling dichotomous data since it is a robust indicator that does not demand that variables are normally distributed (e.g., Brown, 2015; DiStefano et al., 2019). Model fits were verified using comparative fit index (CFI), normed fit index (NFI), Tucker Lewis index (TLI) and root mean square error of approximation (RMSEA). The CFI ≥ 0.90, NFI ≥ 0.90, TLI ≥ 0.90 and RMSEA < 0.10 show a non-rejection of the models, while CFI ≥ 0.95, NFI ≥ 0.95, TLI ≥ 0.95 and RMSEA < 0.06 are indicators of a good model fit (Putnick & Bornstein, 2016; Schumacker & Lomax, 2018).

For each not rejected model, the standardized factor loadings, factor correlations, and reliability of the scores were calculated. McDonald's omega (Flora, 2020) was used to calculate reliability. Although there is no consensus in the literature on the minimum acceptable values of the omega, in this study we considered the criteria of Reise et al. (2013) which establishes a minimum value of .50 and preferred value of .75. In this study we only use Mcdonald's omega (Ω) as an indicator of reliability and not Cronbach's alpha, considering the disadvantages widely discussed in the literature about the latter (e.g., Flora, 2020; McNeish, 2018).

The bifactor model tests the variance of any specific latent variable at the presence of the general factor, to the extent that all latent variables are orthogonalized. Therefore, if any latent variable had variance zero, then the bifactor model would be run again, without the presence of this latent variable, and so on, until only the latent variables with positive variance remained in the model. To define the model with the best fit, not only the highest CFI, NFI, TLI and lowest RMSEA values were considered, but also the factor loadings, variances, and acceptable reliability of the model's factors. All analyses were performed in R software version 3.6.2, using the packages semTools, version 0.5-4 (Jorgensen et al., 2021) and lavaan, version 0.6-7 (Rosseel et al., 2020).

## Results

*Descriptive Analysis of the Items*

Table 1 shows the percentage of correct answers for the items. There is a relatively similar distribution in each of the ranges for the planning, monitoring, and judgment items. Most items in these abilities are in the medium (hits between 41% and 60%) and difficult (hits between 21% and 40%) categories, but there are items in all of the hit ranges with the exception of the range between 0 and 20% hits in the judgment ability.

*Table 1. Percentage of Correct Answers for the Items*

| % Hits | Planning | Monitoring | Judgment |
|---|---|---|---|
| 81 – 100 | 1, 16 | 1, 16 | 1, 16 |
| 61 – 80 | 2, 4, 13 | 2, 13 | 2, 4, 13, |
| 41 – 60 | 3, 6, 9, 10, 11, 14 | 3, 4, 6, 9, 10, 11 | 3, 6, 9, 10, 11, 14, 18 |
| 21 – 40 | 5, 8, 12, 15, 18 | 5, 8, 12, 14, 18 | 5, 7, 8, 12, 15, 17 |
| 0 – 20 | 7, 17 | 7, 15, 17 | - |

*Item Confirmatory Factor Analysis and Reliability*

The results of the indices of fit of the tested models are presented in Table 2. From the models analyzed, only the three-factor uncorrelated model B showed indices of fit below acceptable values (CFI, NFI & TLI < .90; RMSEA > .08) and it was therefore rejected.

*Table 2. Indices of Fit of the Tested Models*

| Model | $\chi^2$ (df) | *p* | CFI | NFI | TLI | RMSEA (CI 90%) |
|---|---|---|---|---|---|---|
| A. Unidimensional | 4758.15 (1359) | .000 | .925 | .899 | .921 | .062 (.060 - .064) |
| B. Uncorrelated factors | 11455.44 (1359) | .000 | .778 | .756 | .766 | .107 (.105 - .108) |
| C. Correlated factors | 2185.27 (1356) | .000 | .982 | .953 | .981 | .031 (.028 - .033) |
| D. Hierarchical (g and 3 specific) | 2185.27 (1356) | .000 | .982 | .953 | .981 | .031 (.028 - .033) |
| E. Bifactor (g and 3 specific) | 1470.07 (1305) | .000 | .996 | .969 | .996 | .014 (.009 - .018) |
| *Variants E- Bifactor Model* | | | | | | |
| E.1. Bifactor (g and 2 specific) | 1591.60 (1323) | .000 | .994 | .966 | .994 | .018 (.014 - .021) |
| E.2. Bifactor (g and 1 specific) | 1727.78 (1341) | .000 | .992 | .963 | .991 | .021 (.018 - .024) |

$\chi^2$ = chi-square; df = degrees of freedom; CFI = comparative fit index; NFI = normed fit index; TLI = Tucker Lewis index; RMSEA = root mean square error of approximation; CI = confidence interval.

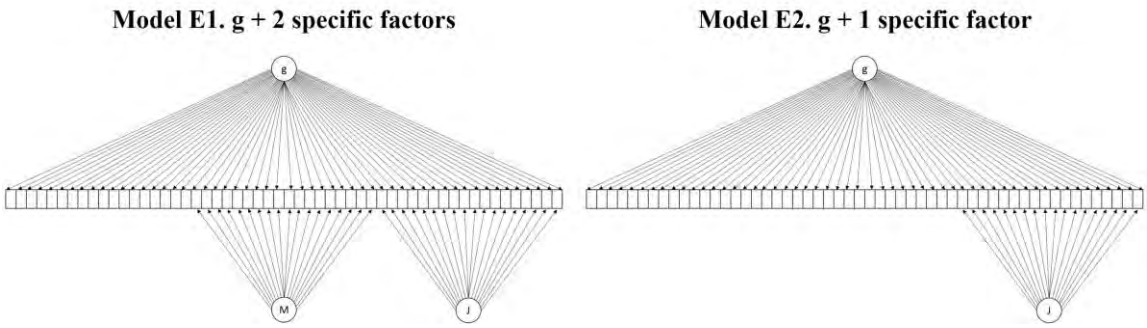
Model A considers a unidimensional structure of the items, it showed acceptable values of CFI (> .90), TLI (> .90) and RMSEA (< .08), but a non-acceptable value of NFI (< .90). The standardized factor loadings of this model ranged from -.049 to .850 (M = .486; SD = .198). The reliability of the latent variable was Ω = .758.

A good fit to the data was obtained by the three-factor correlated in Model C (CFI, NFI & TLI > .95; RMSEA < .05). Planning showed factorial loadings between .326 and .797 (M = .577; SD = .126) with a reliability of Ω = .797. Monitoring showed loadings between .303 and .879 (M = .627; SD = .152) and Ω = .753. Judgment, on the other hand, showed loadings between .471 and .826 (M = .680; SD = .103) and Ω = .753. The correlations of the latent variables in the model were .912 (p < .0009) between planning and monitoring, .278 (p < .0009) between planning and judgment, and .158 (p = .020) between monitoring and judgment.

Model D (hierarchical) also showed a good fit to the data (CFI, NFI & TLI > .95; RMSEA < .05). Standardized factor loadings ranged from .303 to .879 (M = .628; SD = .133) for the first-level latent variables. The overall second-level factor had factor loadings of 1.266 (p = .020) on planning, .720 (p = .000) on monitoring, and .219 (p = .022) on judgment. The variance of the specific planning factor has negative and therefore zero variance ($S^2$ = -0.173). The reliability indices were the same as model C for each specific latent variable.

Model E (bifactor) was the model with the best fit (CFI = .996; NFI = .969; TLI = .996; RMSEA = .014). However, this model presented some problems linked to two specific factors. The model showed very low factor loadings of the specific factors of planning and monitoring and the variance of these factors was zero at the presence of the general factor of cognition regulation. The standardized factorial loadings of the specific latent variables ranged from -.480 to .287 (M = -.037; SD = .246) for planning, from -.265 to .572 (M = .247; SD = .181) for monitoring and from .579 to .760 (M = .670; SD = .048) for judgment. The factor loadings of cognition regulation ranged from -.193 to .849 (M = .432; SD = .250). With regard to reliability, the results indicated some problems. The general factor of cognition regulation and judgment showed acceptable values (Ω = .703 and .693, respectively). However, the reliability scores were well below the minimum expected for both planning (Ω = .015) and monitoring (Ω = .049).

Following the results of Model E, two variants of this model were analyzed, removing the specific factors of planning and monitoring. In the first variant, Model E.1, only planning was removed, that is, a general factor of cognition regulation and two specific factors of monitoring and judgment were considered. In the second variant, both planning and monitoring factors were removed, therefore, the model included only a general factor and the single judgment factor (see Figure 2).




*Note: g = Regulation of Cognition; M = Monitoring; J = Judgement.*

*Figure 2. Variants of the Bifactor Model*

The fit indices for Models E.1 and E.2 are presented in Table 2. The results indicate good indices of fit for both models (CFI, NFI & TLI > .95; RMSEA < .05). In Model E.1 the standardized factor loadings of monitoring ranged from -.568 to .475 (M = .187; SD = .248) and for judgment showed ranges from .579 to .760 (M = .670; SD = .048). For the general factor, that is, cognition regulation, the factor loadings ranged from -.193 to .849 (M = .432; SD = .250). The factor reliability was Ω = .704 for the cognition regulation factor, Ω = .051 for monitoring and Ω = .696 for judgment. The results of this model indicate that the specific factor of monitoring continued to show factor loadings and reliability far below acceptable values.

Finally, the results of Model E.2 show factor loadings between -.093 and .810 (M = .547; SD = .244) for the general cognition regulation factor and between .579 to .760 (M = .670; SD = .048) for the judgment factor (see Table 3). The judgment items loaded well only for the judgment factor (λ > .30), having lower loadings on the cognition regulation factor (λ < .30). Reliability proved acceptable, with values of Ω = .701 and Ω = .699 for regulation of cognition and judgment, respectively.

*Table 3. Standardized Factorial Loads of Model E2*

| Items | Factors | | Items | Factors | | Items | Factors | |
|---|---|---|---|---|---|---|---|---|
| | g | J | | g | J | | g | J |
| P1 | **.525*** | - | M1 | **.561*** | - | J1 | .217* | **.677*** |
| P2 | **.778*** | - | M2 | **.604*** | - | J2 | .099 | **.717*** |
| P3 | **.387*** | - | M3 | **.382*** | - | J3 | .044 | **.579*** |
| P4 | **.321*** | - | M4 | **.395*** | - | J4 | **.338*** | **.606*** |
| P5 | **.517*** | - | M5 | **.756*** | - | J5 | .125* | **.620*** |
| P6 | **.431*** | - | M6 | **.516*** | - | J6 | .021 | **.610*** |
| P7 | **.564*** | - | M7 | **.626*** | - | J7 | .188* | **.660*** |
| P8 | **.749*** | - | M8 | **.684*** | - | J8 | **.335*** | **.720*** |
| P9 | **.581*** | - | M9 | **.485*** | - | J9 | .238* | **.640*** |
| P10 | **.575*** | - | M10 | **.544*** | - | J10 | .205* | **.661*** |
| P11 | **.624*** | - | M11 | **.497*** | - | J11 | .133* | **.651*** |
| P12 | **.528*** | - | M12 | **.720*** | - | J12 | -.093 | **.678*** |
| P13 | **.706*** | - | M13 | **.588*** | - | J13 | **.330*** | **.726*** |
| P14 | **.470*** | - | M14 | **.790*** | - | J14 | -.069 | **.700*** |
| P15 | **.694*** | - | M15 | **.863*** | - | J15 | .239* | **.702*** |
| P16 | **.648*** | - | M16 | **.592*** | - | J16 | **.361*** | **.682*** |
| P17 | **.507*** | - | M17 | **.810*** | - | J17 | .138* | **.695*** |
| P18 | **.572*** | - | M18 | **.715*** | - | J18 | .072 | **.760*** |

g = Regulation of cognition; P = Planning; M = Monitoring; J = Judgment; * = p < .05; bold values indicate loadings > .30.

## Discussion

The purpose of this study was to assess the structural validity of the Meta-Text by analyzing the dimensionality and reliability of the test. Seven models were tested (five initial models and two additional models). The models were analyzed according to their indices of fit, factor loadings, reliability, and inter-factor correlations. According to the analyses performed, the bifactor model E2, containing a general cognition regulation factor and a judgment-specific factor, was evaluated as the model with the best characteristics, considering its indices of fit (CFI = .992; NFI = .963; TLI = .991; RMSEA = .021) and acceptable factor reliability (Ω = .701 and .699).

In spite of Model A showing an acceptable fit and Models C, D, and E showing good fits, they highlighted some important aspects worth discussing. Model A assumes the unidimensional assumption of metacognition consistent with Immekus and Imbrie's (2008) postulate. Despite showing an acceptable fit, the latent variable of the one-dimensional model loads well only on planning and monitoring items (λ > .30), while the factor loadings of the latent variable are very low on the vast majority of the judgment items (λ < .30 = J2, J3, J5, J6, J7, J11, J12, J14, J17, and J18). This result highlights the fact that the latent variable of the one-dimensional model does not explain the variance of the judgment items, implying that another latent variable, at least, would be needed in the model. Therefore, the cognition regulation latent variable does not alone explain the variance of the Meta-Text items, indicating that the cognition regulation domain probably involves both this broad ability as some more specific ability.

Models C and D, on the other hand, presented problems linked to the correlations between the factors and the loadings of factor g on the specific factors, respectively. In Model C, the specific factors of planning and monitoring showed a very high correlation (r = .912), indicating that both factors share a variance of 83.174%, providing hints about a strong possibility that both latent variables may in fact be a single construct rather than distinct constructs. In the case of model D, the factor loadings of the general factor on the factors planning and monitoring were high (λ = 1.266 and .720), with a factor loading of 1.266 indicating that there are problems in the model and most likely the planning factor does not hold. The results of the two-factor models E and E1 also reinforce the evidence to refute the planning factor and also indicate the need to refute the monitoring factor.

The evidence from models C, D, E, and E1 can be analyzed from the perspective of two hypotheses. The first hypothesis indicates that it is possible that the planning and monitoring constructs are actually just regulation of cognition and do not differ from this general domain. The meta-analysis by Craig et al. (2020) presents evidence on the considerable link between planning and monitoring, showing a correlation of .63 (95% confidence interval = .46-.81). However, the data presented in this meta-analysis primarily involve self-report measures. It is plausible that when these abilities are measured by performance-based instruments, as in the case of the Meta-Text, they show the actual relationship between these abilities, indicating that in reality they are not two distinguished specific processes (Rose et al., 2015).

The second hypothesis concerns the difficulty of developing tasks and items that separate the planning and monitoring factors. Important challenges may be generated while evaluating specific metacognitive abilities, mainly due to the difficulty of separating them from the performance of the task itself and other cognitive and metacognitive components involved in it (e.g., Li et al., 2015; Rose et al., 2015). Moreover, in Meta-Text, the pairs of planning and monitoring items are linked to the same question, so the items have local dependency which is considered in all the models analyzed. In this sense, it is possible that the characteristics of the test may influence the results. A suggestion for empirical verification of this possibility would be the development of tests in which planning and monitoring are not measured in questions that link the measures of both constructs.

In the light of the issues presented in models A, C, D, E and E1, model E2 is a bifactor model in which the presence of specific factors for monitoring and planning is eliminated, although the judgment factor is kept. In model E2, both the general factor and the specific factor showed acceptable reliability. The evidence from Model E2 brings conceptual implications for metacognition theory. Since the judgment items were well loaded only by the judgment factor and show a very weak association with the general factor of cognition regulation (see Table 3), it is suggested that the judgment construct may actually be a component of the metacognitive knowledge dimension. To date, metacognitive theory has preferentially assumed that judgment is a component of cognition regulation. Our evidence opens up the possibility that this assumption of theory is mistaken.

A plausible explanation for those novel findings in our study is the fact that the Meta-Text test is performance-based, bringing new evidence as a function of a more suitable methodology for the measurement of metacognition. Our results seem promising, as initial evidence from studies in the field of cognitive neuroscience is consistent with our results. Despite sharing similar regions of the prefrontal cortex, there are indications that judgment processes may have different mechanisms of functioning, in contrast to more general metacognitive abilities linked to the regulation domain of cognition (e.g., Fleur et al., 2021; Morales et al., 2018). However, the neurocognitive architecture underlying metacognition still needs to be further explored in future research.

The implementation of bifactor models in the field of measuring metacognition is relatively recent. Some studies analyzing this type of models have found good fit rates, nevertheless, the analyses have been performed exclusively on self-report instruments (Fergus & Bardeen, 2019; Ning, 2019; Zhao et al., 2019). Furthermore, the bifactorial

frameworks tested have included as specific factors, metacognitive beliefs (Fergus & Bardeen, 2019), broad domains of metacognition, i.e., cognition and cognition regulation (Ning, 2019), and domain-dependent factors, i.e., reading and mathematics (Zhao et al., 2019). To the best of our knowledge, our study is the first to test bifactor models in the metacognitive field using items based on respondent performance. This type of bifactor model had not yet been tested in the field.

## Conclusion

This study presents the first evidence on the structural validity of the Meta-Text. The results support a bifactor structure of the test that includes a general cognition regulation factor and a judgment specific factor. The shown evidence offers three important implications for the field of study of metacognition.

The first implication concerns the use of performance tests to measure metacognition. The use of performance tests allows one to deal with the respondent and confirmatory biases contained in self-report and think-aloud instruments.

The second implication is related to the theorization of metacognitive domains and abilities. Traditionally, planning, monitoring, and judgment have been linked to the domain of regulation of cognition. However, the results provide evidence that planning and monitoring are only regulation of cognition and not specific processes and that judgment is possibly a component of metacognitive knowledge and not of regulation of cognition.

The third implication concerns the importance of implementing bifactor models in psychometric studies of metacognition measures. By using these models one can to verify whether the specific factors remain valid in the presence of a general factor, separating the variance attributable to each factor. To the best of our knowledge, bifactor models of metacognition are only beginning to be addressed in recent research, so a broader understanding and application of these models in the metacognitive field is still needed. The present study analyzes and distinguishes different factor structures, incorporating a bifactorial structure which brings promising evidence, from performance-based test data.

## Recommendations

As a research agenda, it is important to conduct studies to test the structural validity of the Meta-Text in other higher education samples with different sociodemographic and cultural characteristics. Furthermore, in order to test whether the factor structure of the test is replicable in different population groups, it is necessary to perform invariance analyses of the test scores according to gender, course, educational level, nationality, or other variables of sociodemographic and educational interest.

Considering the complexity in measuring metacognition and taking into account that there is still no consensus in the literature about the best instruments to measure it, the use of multi-method research designs is essential. Conducting studies involving different metacognitive measures (e.g., Think Aloud protocols and performance-based tests) represents an important field of study. These types of studies are relevant for obtaining a more comprehensive and accurate picture of students' metacognitive abilities (Gascoine et al., 2017).

Finally, in order to add new evidence on the validity of the Meta-Text and to make it available for use in psychopedagogical diagnosis and intervention processes, it is important to develop future investigations to test the external validity of the test, especially its link with educational outcomes. Furthermore, it is indicated that the other test of the Meta-Performance Battery should also be evaluated in terms of its validity, and that both tests could be evaluated together.

## Limitations

Despite the contributions of this study to the field of metacognition, some limitations need to be pointed out. The sample used in this study was selected in a non-probabilistic way by convenience, so generalization of results is not possible. Furthermore, Meta-Text assesses metacognition solely in the domain of reading and comprehension of texts, so the findings of this study may be domain dependent. Finally, although Google Forms is one of the most widely used platforms in online data collection in different fields of knowledge (Mondal et al., 2019), more robust software specialized in measuring cognitive constructs should be used allowing the capture of other information, such as the subject's interactions with the test or response time on each item and on the test in general.

### Authorship Contribution Statement

Castillo-Diaz: Conceptualization, design, data acquisition, data analysis, interpretation, drafting manuscript, critical revision of manuscript, final approval. Gomes: Conceptualization, design, data analysis, interpretation, drafting manuscript, critical revision of manuscript, supervision, final approval. Jelihovschi: Conceptualization, design, interpretation, drafting manuscript, writing, critical revision of manuscript, final approval.

**References**

Abernethy, M. (2015). Self-reports and observer reports as data generation methods: An assessment of issues of both methods. *Universal Journal of Psychology*, *3*(1), 22–27. https://doi.org/10.13189/ujp.2015.030104

Akturk, A., & Sahin, I. (2011). Literature review on metacognition and its measurement. *Procedia Social and Behavioral Sciences*, *15*, 3731–3736. https://doi.org/10.1016/j.sbspro.2011.04.364

Amin, A., Corebima, A., Zubaidah, S., & Mahanal, S. (2020). The correlation between metacognitive skills and critical thinking skills at the implementation of four different learning strategies in animal physiology lectures. *European Journal of Educational Research, 9*(1), 143–163. https://doi.org/10.12973/eu-jer.9.1.143

Azevedo, R. (2020). Reflections on the field of metacognition: Issues, challenges, and opportunities. *Metacognition and Learning*, *15*(2), 91–98. https://doi.org/10.1007/s11409-020-09231-x

Brown, T. A. (2015). *Confirmatory factor analysis for applied research* (2nd ed.). The Guilford Press.

Castillo-Diaz, M. A., & Gomes, C. M. A. (2021). Presenting the Meta-Performance Test, a metacognitive battery based on performance. *International Journal of Educational Methodology, 7*(2), 289-303. https://doi.org/gjwgpv

Castillo-Diaz, M. A., & Gomes, C. M. A. (2022). Monitoring and intelligence as predictors of a standardized measure of general and specific higher education achievement. *Trends in Psychology*. Advance online publication. https://doi.org/10.1007/s43076-022-00160-z

Craig, K., Hale, D., Grainger, C., & Stewart, M. E. (2020). Evaluating metacognitive self-reports: Systematic reviews of the value of self-report in metacognitive research. *Metacognition and Learning, 15*(2), 155-213. https://doi.org/10.1007/s11409-020-09222-y

Cromley, J. G., & Kunze, A. J. (2020). Metacognition in education: Translational research. *Translational Issues in Psychological Science, 6*(1), 15-20. https://doi.org/10.1037/tps0000218

Dent, A. L., & Koenka, A. C. (2016). The relation between self-regulated learning and academic achievement across childhood and adolescence: A meta-analysis. *Educational Psychology Review*, *28*(3), 425–474. https://doi.org/10.1007/s10648-015-9320-8

Desoete, A., Roeyers, H., & Buysse, A. (2001). Metacognition and mathematical problem solving in grade 3. *Journal of Learning Disabilities*, *34*(5), 435–447. https://doi.org/10.1177/002221940103400505

DiStefano, C., McDaniel, H. L., Zhang, L., Shi, D., & Jiang, Z. (2019). Fitting large factor analysis models with ordinal data. *Educational and Psychological Measurement*, *79*(3), 417–436. https://doi.org/10.1177/0013164418818242

Donker, A. S., de Boer, H., Kostons, D., van Dignath Ewijk, C. C., & van der Werf, M. (2014). Effectiveness of learning strategy instruction on academic performance: A meta-analysis. *Educational Research Review*, *11*, 1–26. https://doi.org/10.1016/j.edurev.2013.11.002

Fergus, T. A., & Bardeen, J. R. (2019). The Metacognitions Questionnaire-30: An examination of a bifactor model and measurement invariance among men and women in a community sample. *Assessment*, *26*(2), 223–234. https://doi.org/10.1177/1073191116685807

Filippi, R., Ceccolini, A., Periche-Tomas, E., & Bright, P. (2020). Developmental trajectories of metacognitive processing and executive function from childhood to older age. *Quarterly Journal of Experimental Psychology*, *73*(11), 1757–1773. https://doi.org/10.1177/1747021820931096

Fleur, D. S., Bredeweg, B., & van den Bos, W. (2021). Metacognition: Ideas and insights from neuro- and educational sciences. *NPJ Science of Learning*, *6*(1), Article 13. https://doi.org/10.1038/s41539-021-00089-5

Flora, D. B. (2020). Your coefficient alpha is probably wrong, but which coefficient omega is right? A tutorial on using r to obtain better reliability estimates. *Advances in Methods and Practices in Psychological Science*, *3*(4), 484–501. https://doi.org/10.1080/19312458.2020.1718629

Gascoine, L., Higgins, S., & Wall, K. (2017). The assessment of metacognition in children aged 4–16 years: *A systematic review. Review of Education, 5*(1), 3–57. https://doi.org/10.1002/rev3.3077

Golino, H. F., & Gomes, C. M. A. (2011). Preliminary internal validity evidences of two Brazilian Metacognitive Tests. *International Journal of Testing, 26*, 11-12. https://www.intestcom.org/files/ti26.pdf

Gomes, C. M. A. (2021, September 1-3). *Presentation of a methodology for creating metacognitive tests* [Paper presentation]. International Galician-Portuguese Congress of Psychopedagogy, University of Minho, Braga, Portugal. https://doi.org/10.13140/RG.2.2.33129.62569

Gomes, C. M. A., Araujo, J. D., & Castillo-Diaz, M. A. (2021). Testing the invariance of the Metacognitive Monitoring Test. *Psico-USF*, *26*(4), 685–696. https://doi.org/10.1590/1413-82712021260407

Gomes, C. M. A., & Golino, H. F. (2014). Self-reports on students' learning processes are academic metacognitive knowledge. *Psychology: Reflection and Criticism/ Psicologia: Reflexão e Crítica, 27*(3), 472-480. https://doi.org/10.1590/1678-7153.201427307

Gomes, C. M. A., Golino, H. F., & Menezes, I. G. (2014). Predicting school achievement rather than intelligence: Does metacognition matter? *Psychology*, *5*, 1095–1110. https://doi.org/10.4236/psych.2014.59122

Gomes, C. M. A., Linhares, I., Jelihovschi, E., & Rodrigues, M. (2021). Introducing rationality and content validity of slat-thinking. *International Journal of Development Research*, *11*(1), 43264–43272. https://bit.ly/3fSxBMM

Gomes, C. M. A., & Nascimento, D. (2021). Presenting slat-thinking second version and its content validity. *International Journal of Development Research*, *11*(3), 45590–45596. https://bit.ly/3rFwByt

Gomes, C. M. A., Quadros, J. S., Araujo, J., & Jelihovschi, E. G. (2020). Measuring students' learning approaches through achievement: Structural validity of SLAT-Thinking. *Psychology Studies/ Estudos de Psicologia, 25*(1), 33-43. https://bit.ly/3RCHFHf

Greene, J. A., Deekens, V. M., Copeland, D. Z., & Yu, S. (2018). Capturing and modeling self-regulated learning using think-aloud protocols. In D. H. Schunk & J. A. Greene (Eds.), *Handbook of self-regulation of learning and performance* (2nd ed., pp. 323–337). Routledge. https://doi.org/10.4324/9781315697048-21

Hu, J., & Gao, X. (2017). Using think-aloud protocol in self-regulated reading research. *Educational Research Review*, *22*, 181–193. https://doi.org/10.1016/j.edurev.2017.09.004

Immekus, J. C., & Imbrie, P. K. (2008). Dimensionality assessment using the full-information item bifactor analysis for graded response data: An illustration with the State Metacognitive Inventory. *Educational and Psychological Measurement*, *68*(4), 695–709. https://doi.org/10.1177/0013164407313366

Jansen, R. S., van Leeuwen, A., Janssen, J., Jak, S., & Kester, L. (2019). Self-regulated learning partially mediates the effect of self-regulated learning interventions on achievement in higher education: A meta-analysis. *Educational Research Review*, *28*, Article 100292. https://doi.org/10.1016/j.edurev.2019.100292

Jia, X., Li, W., & Cao, L. (2019). The role of metacognitive components in creative thinking. *Frontiers in Psychology, 10*, Article 2404. https://doi.org/10.3389/fpsyg.2019.02404

Jorgensen, T. D., Pornprasertmanit, S., Schoemann, A. M., & Rosseel, Y. (2021). *semTools: Useful tools for structural equation modeling. R package* (version 0.5-4) [Computer software]. https://bit.ly/3s5bZjd

Li, J., Zhang, B., Du, H., Zhu, Z., & Li, Y. M. (2015). Metacognitive planning: Development and validation of an online measure. *Psychological Assessment, 27*(1), 260-271. https://doi.org/10.1037/pas0000019

McNeish, D. (2018). Thanks coefficient alpha, we'll take it from here. *Psychological Methods*, *23*(3), 412–433. https://doi.org/10.1037/met0000144

Mondal, H., Mondal, S., Ghosal, T., & Mondal, S. (2019). Using Google Forms for medical survey: A technical note. *International Journal of Clinical and Experimental Physiology*, *5*(4), 216–218. https://doi.org/10.5530/ijcep.2018.5.4.26

Morales, J., Lau, H., & Fleming, S. M. (2018). Domain-general and domain-specific patterns of activity supporting metacognition in human prefrontal cortex. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, *38*(14), 3534–3546. https://doi.org/10.1523/JNEUROSCI.2360-17.2018

Muijs, D., & Bokhove, C. (2020). *Metacognition and self-Regulation: Evidence review*. Education Endowment Foundation. https://bit.ly/3VaqeAv

Neuenhaus, N., Artelt, C., Lingel, K., & Schneider, W. (2011). Fifth graders metacognitive knowledge: General or domain-specific? *European Journal of Psychology of Education*, *26*(2), 163–178. https://doi.org/czv78g

Ning, H. K. (2019). The bifactor model of the Junior Metacognitive Awareness Inventory (Jr. MAI). *Current Psychology*, *38*(2), 367–375. https://doi.org/10.1007/s12144-017-9619-3

Norman, E., Pfuhl, G., Sæle, R. G., Svartdal, F., Låg, T., & Dahl, T. I. (2019). Metacognition in psychology. *Review of General Psychology*, *23*(4), 403–424. https://doi.org/10.1177/1089268019883821

Ohtani, K., & Hisasaka, T. (2018). Beyond intelligence: A meta-analytic review of the relationship among metacognition, intelligence, and academic performance. *Metacognition and Learning, 13*(2), 179–212. https://doi.org/10.1007/s11409-018-9183-8

Oliveira, A., & Nascimento, E. (2014). Construção de uma escala para avaliação do planejamento cognitivo [Construction of a cognitive planning assessment scale]. *Psychology: Reflection and Criticism/ Psicologia: Reflexão e Crítica, 27*(2), 209-218. https://doi.org/10.1590/1678-7153.201427201

Panadero, E. (2017). A review of self-regulated learning: Six models and four directions for research. *Frontiers in Psychology*, *8*, Article 422. https://doi.org/10.3389/fpsyg.2017.00422

Pires, A. A. M., & Gomes, C. M. A. (2017). Three mistaken procedures in the elaboration of school exams: Explicitness and discussion. *PONTE International Scientific Researches Journal, 73*(3), 1-14. https://doi.org/10.21506/j.ponte.2017.3.1

Pires, A. A. M., & Gomes, C. M. A. (2018). Proposing a method to create metacognitive school exams. *European Journal of Education Studies, 5*(8), 119-142. https://doi.org/10.5281/zenodo.2313538

Preiss, D., Ibaceta, M., Ortiz, D., Carvacho, H., & Grau, V. (2019). An exploratory study on mind wandering, metacognition, and verbal creativity in Chilean high school students. *Frontiers in Psychology, 10*, Article 1118. https://doi.org/10.3389/fpsyg.2019.01118

Putnick, D. L., & Bornstein, M. H. (2016). Measurement invariance conventions and reporting: The state of the art and future directions for psychological research. *Developmental Review*, *41*, 71-90. https://doi.org/10.1016/j.dr.2016.06.004

Reise, S. P. (2012). The rediscovery of bifactor measurement models. *Multivariate behavioral research*, *47*(5), 667–696. https://doi.org/10.1080/00273171.2012.715555

Reise, S. P., Bonifay, W. E., & Haviland, M. G. (2013). Scoring and modeling psychological measures in the presence of multidimensionality. *Journal of Personality Assessment, 95,* 129–140. https://doi.org/gfrkkf

Roebers, C. M. (2017). Executive function and metacognition: Towards a unifying framework of cognitive self-regulation*. Developmental Review, 45,* 31–51. https://doi.org/10.1016/j.dr.2017.04.001

Rose, N. S., Luo, L., Bialystok, E., Hering, A., Lau, K., & Craik, F. I. M. (2015). Cognitive processes in the Breakfast Task: Planning and monitoring. *Canadian Journal of Experimental Psychology/ Revue Canadienne De Psychologie Experimentale*, *69*(3), 252–263. https://doi.org/10.1037/cep0000054

Rosseel, Y., Jorgensen, T. D., Rockwood, N., Oberski, D., Byrnes, J., Vanbrabant, L., Savalei, V., Merkle, E., Hallquist, M., Rhemtulla, M., Katsikatsou, M., Barendse, M., Scharf, F., & Du, H. (2020). *Lavaan: Latent Variable Analysis. R package (version 0.6-7)* [Computer software]. https://bit.ly/3gmzbqR

Saenz, G. D., Geraci, L., & Tirso, R. (2019). Improving metacognition: A comparison of interventions. *Applied Cognitive Psychology*, *33*(5), 918–929. https://doi.org/10.1002/acp.3556

Schraw, G. (2009). Measuring metacognitive judgments. In D. J. Hacker, J. Dunlosky, & A. C. Graesser (Eds.), *Handbook of metacognition in education* (pp. 415-429). Routledge.

Schumacker, R., & Lomax, R. (2018). *A beginner's guide to structural equation modeling* (4th ed.). Routledge.

Schunk, D. H., & Greene, J. A. (Eds.). (2018). *Handbook of self-regulation of learning and performance* (2nd ed.). Routledge. https://doi.org/10.4324/9781315697048

Silva, C., & Iturra, C. (2021). A conceptual proposal and operational definitions of the cognitive processes of complex thinking. *Thinking Skills and Creativity*, *39*, Article 100794. https://doi.org/10.1016/j.tsc.2021.100794

Van der Stel, M., & Veenman, M. (2008). Relation between intellectual ability and metacognitive skillfulness as predictors of learning performance of young students performing tasks in different domains. *Learning and Individual Differences, 18*(1), 128–134. https://doi.org/10.1016/j.lindif.2007.08.003

Veenman, M., & Van Cleef, D. (2018). Measuring metacognitive skills for mathematics: Students' self-reports versus on-line assessment methods. *ZDM, 51*(4), 691-701. https://doi.org/10.1007/s11858-018-1006-5

Wetzel, E., Böhnke, J. R., & Brown, A. (2016). Response biases. In F. T. L. Leong, D. Bartram, F. M. Cheung, K. F. Geisinger, & D. Iliescu (Eds.), *The ITC international handbook of testing and assessment* (pp. 349-363). Oxford University Press. https://doi.org/ghm7qr

Wolcott, M. D., & Lobczowski, N. G. (2021). Using cognitive interviews and think-aloud protocols to understand thought processes. *Currents in Pharmacy Teaching and Learning, 13*(2), 181-188. https://doi.org/jg8g

Zhao, N., Teng, X., Li, W., Li, Y., Wang, S., Wen, H., & Yi, M. (2019). A path model for metacognition and its relation to problem-solving strategies and achievement for different tasks. *ZDM*, *51*(4), 641–653. https://doi.org/10.1007/s11858-019-01067-3