Testing the Invariance of the Metacognitive Monitoring Test

Cristiano Mauro Assis Gomes[1]
Jhonys de Araujo[1]
Marcio Alexander Castillo-Díaz[1,2]

[1]*Universidade Federal de Minas Gerais, Belo Horizonte, Minas Gerais, Brasil*
[2]*Universidad Nacional Autónoma de Honduras, Tegucigalpa, Honduras*

Abstract
Metacognition is predominantly measured by the self-report and think-aloud methods. This is problematic since they produce considerable both respondent and confirmatory biases, which implies damage to the measurement. The Metacognitive Monitoring Test (MMT) was created to evaluate metacognition through performance and eliminate the aforementioned biases. There is evidence of MMT convergent, divergent, structural, predictive and incremental validity. This article expands the validity studies about the MMT by analyzing the configural, metric and scalar invariance of MMT across sex, nationality, and educational level variables. The sample is composed of Brazilian and Honduran subjects, as well as 6st to 12st grades and higher education students. Results indicate configural, metric and scalar invariance for the sex variable, as well as configural invariance and metric and scalar partial invariance for nationality and educational level. It is concluded that the MMT allows comparing means of the latent variable measured in the analyzed groups.
Keywords: metacognition; self-report; test validity; performance tests.

Investigando a Invariância do Teste de Monitoramento Metacognitivo

Resumo
Há uma hegemonia dos métodos de autorrelato e *thinkaloud* para avaliar metacognição. Isso é problemático, pois eles geram substanciais vieses do respondente e confirmatório, trazendo prejuízo à medida. O Teste de Monitoramento Metacognitivo (TMC) foi criado para avaliar a metacognição mediante o desempenho e eliminar os vieses supramencionados. Há evidências de validade convergente, divergente, estrutural, preditiva e incremental do TMC. Este artigo amplia os estudos de validade e analisa a invariância configural, métrica e escalar do TMC, em relação ao sexo, nacionalidade, e nível educacional. A amostra do estudo é composta por brasileiros e hondurenhos, e estudantes da 6ª à 12ª séries da educação básicae ensino superior. Os resultados indicam invariância configural, métrica e escalar para a variável sexo, assim como invariância configural e invariância parcial métrica e escalar para nacionalidade e nível educacional. Conclui-se que o TMC permite comparar médias da variável latente mensurada nos grupos analisados.
Palavras-chave: metacognição; autorrelato; validade do teste; testes de desempenho.

Investigando la Invarianza de la Prueba de Monitoreo Metacognitivo

Resumen
Hay un predominio de métodos de auto-reporte e *think aloud* para evaluar la metacognición. Esto es problemático, ya que dichos métodos producen un considerable sesgo de respuesta y confirmación, lo que implica un perjuicio en la medición. La Prueba de Monitoreo Metacognitivo (PMM) fue diseñada para evaluar la metacognición por medio del desempeño y eliminar los sesgos mencionados anteriormente. Hay evidencia de validez convergente, divergente, estructural, predictiva e incremental de la PMM. Este artículo amplía los estudios de validez sobre la PMM, analizando la invarianza configural, métrica y escalar de la PMM con respecto a las variables sexo, nacionalidad y nivel educativo. La muestra del estudio está formada por brasileños y hondureños, y estudiantes de 6º a 12º grado de educación básica y educación superior. Los resultados indican invarianza configural, métrica y escalar para la variable sexo así como invarianza configural e invarianza parcial escalar y métrica para nacionalidad y nivel educativo. Se concluye que la PMM permite comparar medias de la variable latente medida en los grupos analizados.
Palabras clave: metacognición; autorreporte; validación de test; test de desempeño.

Metacognition is a construct investigated by different traditions and areas of knowledge, such as neuroscience, psychology, education and linguistics (Bártolo-Ribeiro, Simões, & Almeida, 2015; Efklides & Misailidi, 2010; Fleming & Frith, 2014; Haukås, Bjørke, & Dypedahl, 2018). Each of these areas proposes different models to explain the components of metacognition and its mechanisms, implying the development of different data collection and measurement tools and instruments. Nevertheless, there is consensus in the literature that metacognition refers to the ability to know, monitor and self-regulate one's own cognitive processes, which has an impact on the management and control of learning. Furthermore, there is consensus

http://dx.doi.org/10.1590/1413-82712021260407

that metacognition comprises two different broad dimensions: metacognitive knowledge and metacognitive regulation, where the former concerns the ability to know one's own processes, and the latter involves the mechanisms to control the performance of tasks (Cromley & Kunze, 2020; Peña-Ayala, 2015).

Systematic reviews and meta-analyses by Akturk and Sahin (2011), Gascoine, Higgins and Wall (2017), and Ohtani and Hisasaka (2018) show that the study of metacognition is dominated by the use of self-report questionnaires and think-aloud protocols. However, such prevalence is damaging, as these forms of measurement produce substantial bias, causing harm to the measurement.

Self-report questionnaires produce considerable respondent bias as they require the respondents' accurate self-assessment of their own internal processes (Abernethy, 2015; Pintrich, Wolters, & Baxter, 2000). They also produce other biases (e.g. acquiescence, social desirability) that compromise the measurement (Wetzel, Böhnke, & Brown, 2016). In turn, think-aloud procedures imply considerable confirmation bias, as they require judges to evaluate and classify the respondents' performance and speech. As explained by Das-Smaal (1990, p. 349), "...real-world features, objects, and events can be categorized in countless different ways. Moreover, our perception is highly selective and therefore, already readily biased."

The Metacognitive Monitoring Test (also called Reading Monitoring Test or Read Monitoring Test; Golino & Gomes, 2011; Gomes, Golino, & Menezes, 2014) was created with the purpose of generating a measure of the metacognitive monitoring ability without the respondent and confirmation biases. The test score is generated according to the respondents' own performance in identifying certain errors while they work through a task. Every error identified adds a point to the respondent's score, so that the test does not require judges to evaluate the respondents' performance, as it happens in the think-aloud method. Also, the test does not require respondents to report on their own abilities, as it is the case in self-report questionnaires. Following the review by Gascoine et al. (2017), it was identified that, besides the Metacognitive Monitoring Test (MMT), out of a total of 80 measurement procedures, only two other tools produced metacognition scores based on performance, namely the Metacognitive Skills and Knowledge Assessment (MSA; Desoete, Roeyers, & Buysse, 2001) and the Metacognitive Knowledge Test (Neuenhaus, Artelt, Lingel, & Schneider, 2011).

The Metacognitive Monitoring Test (MMT) shows evidence of structural, convergent, divergent, predictive, and incremental validity for elementary, middle, and high school, as well as higher education students. In Golino and Gomes (2011)'s study, there is evidence of structural validity, since the test items are explained solely by a latent variable (monitoring) and show factor loadings above 0.35. This study also shows evidence of divergent validity, as the test items are explained specifically by the monitoring latent variable, as opposed to the other metacognitive ability, the judgment latent variable. Furthermore, the measure generated shows acceptable reliability, with an alpha value of 0.63. In turn, Gomes et al. (2014) provide evidence of predictive and incremental validity, because they show that monitoring alone accounts for about 20% of the variance in academic performance, producing a higher prediction than that provided by fluid intelligence. On the other hand, Gomes and Golino (2014)'s study indicates evidence of convergent and divergent validity, because, when comparing two models (b and c of their analysis) and their fits, they find evidence that monitoring is a metacognitive ability that is part of the regulation of cognition (which they call *working metacognition*) and is different from metacognitive knowledge (which they call *academic metacognitive knowledge*). Finally, Castillo (2018) shows evidence of structural, predictive, and incremental validity of the Metacognitive Monitoring Test in a Honduran sample of university students, while the previous studies used Brazilian samples. He found that the test items are explained by a latent variable (monitoring), with factor loadings between 0.56 and 0.97, and alpha reliability of 0.73. In addition, he identified that monitoring alone accounts for about 40% of the variance in academic performance, much higher than the prediction provided by general intelligence.

Despite the wide range of validity and reliability evidence, the Metacognitive Monitoring Test has not yet been evaluated in terms of its measurement invariance. This analysis is mandatory to assess whether test scores can be used to compare groups (see the recommendation by the International Test Commission, 2017). Therefore, this study aims to investigate the Metacognitive Monitoring Test configural, metric, and scalar invariance by analyzing the sex (female vs. male), educational level (6^{th} to 12^{th} grades vs. higher education) and nationality (Brazilians vs. Hondurans) variables. Configural invariance analysis allows verifying whether the factorial structure identified in Golino and Gomes

(2011) and Castillo (2018) is present in the groups of the variables analyzed. Just to recall, the factorial structure investigated determines that a latent variable, metacognitive monitoring, explains the variance of the nine test items. On the other hand, the metric invariance analysis allows verifying whether factor loadings are invariant between the groups of the variables analyzed. This analysis is essential, because if the factor loadings of one group differ substantially from those of another group, then the factor scores of these groups cannot be compared (Putnick & Bornstein, 2016). The scalar invariance, in turn, evaluates whether the thresholds of each test item are invariant between the groups of the variables analyzed. As pointed out by Putnick and Bornstein (2016), the factor scores of a test can only be used for group comparison purposes when there is evidence of scalar invariance. This study does not aim to perform residual invariance analysis because, as stated by Bowen and Masa (2015), such analysis is not necessary to ensure the comparability of groups, so it can be discarded when the focus is to investigate the validity of a test to compare certain groups.

Method

Participants

This study has three convenience samples. The first sample was collected in 2008 and consists of 716 students from the 6th to the 12th grades of Brazilian basic education. All students come from a private school in the city of Belo Horizonte, Minas Gerais, Brazil. The mean age of this sample is 13.8 years (SD = 2.1), consisting of 337 (47%) male students and 379 (53%) female students. The distribution of students among grades is well balanced: 6th grade = 94 (13%), 7th h grade = 110 (15%), 8th grade = 99 (14%), 9th grade = 119 (17%), 10th grade = 97 (13.5%), 11th grade = 100 (14%) and 12th grade = 97 (13.5%). The second sample was collected in 2017 and consists of 459 students from the most important public university in Honduras, located in Tegucigalpa. This sample consists of young adults, with a mean age of 18.1 years (SD = 2.5), 284 (62%) female students and 175 (38%) male students. Regarding the branches of knowledge, this sample has 140 (30%) students in biological sciences, 96 (21%) students in exact sciences, 119 (26%) students in economic sciences, 104 (23%) students in social sciences, humanities and arts. The third sample was collected in 2018 and includes 500 students from a public exact sciences and technology university located in Joinville, state of Santa Catarina, Brazil. This sample has a mean age of 22.4 years (SD = 4.8), 184 (37%) female students and 316 (63%) male students. Regarding the branches of knowledge, 67 (13.4%) students belong to teaching degree courses, 377 (75.4%) are engineering students and 56 (11.2%) come from computing courses. The three samples amount to 1,675 participants in total, with a mean age of 17.5 years (4.9). They include 1,216 (73%) Brazilian students and 459 (27%) Honduran students, 828 (49%) male students and 847 (51%) female students, and 959 (57%) higher education students and 716 (43%) basic education students (6th to 12th grade).

Instrument

The Metacognitive Monitoring Test (MMT) evaluates the ability to monitor errors while performing a task (Golino & Gomes, 2011). The rationality of the test is linked to Markman (1977, 1979)'s error detection paradigm. To evaluate the monitoring ability, this paradigm uses texts containing specific errors that must be identified by the respondent.

The Metacognitive Monitoring Test consists of a short text of half an A4 sheet page. The font used is 10pt Verdana and spacing between lines is 1.0 cm. The text includes words that are part of the respondents' current knowledge and vocabulary to avoid that they fail to identify the errors due to lack of vocabulary or previous knowledge (Golino & Gomes, 2011; Gomes et al., 2014). The text contains nine intentionally inserted errors, and the identification of each error represents one point. The maximum number of points is the total score of the test, that is, nine points.

The test was especially designed to avoid some common issues found in tasks that use the error detection paradigm. The test instructions emphasize to the respondent that the text they will read contains errors and that the task is to identify such errors. They ask the respondent to read the text carefully and highlight each passage where he/she thinks there is an error, justifying, in a given field, why he/she marked it. Details on the instructions can be found in Golino and Gomes (2011) and Gomes et al. (2014). The caution pointed out in the instruction aims to avoid that respondents fail to detect errors by assuming that the test text is free from inconsistencies, considering that in the academic context students do not tend to read texts with intentionally produced errors. The respondents that best monitor are expected to identify the highest number of errors.

Procedures for data collection

The data of the three samples come from three different studies. The first sample comes from Gomes et al. (2014)'s study, while the second one comes from Castillo (2018)'s study, so that details about the data collection are reported in those studies. The third sample does not come from any published study and, therefore, the collection procedures will be described herein. Participants in this sample received an e-mail link that redirected them to Survey Monkey, an online platform that was used to run the Metacognitive Monitoring Test and record the participants' responses. There was no time limit to complete the test on the online platform. The three data collections followed the ethical procedures recommended in human research.

Data Analysis

There is no methodological consensus in the world literature on how to perform measurement invariance analysis (Bowen & Masa, 2015). This lack of consensus requires that the researcher explain his methodological choices very carefully. Because of that, the major challenges pointed out by the literature will be presented, in a summarized way, in order to support the methodological choices adopted in this study.

In essence, measurement invariance analysis compares and evaluates the fit of models that represent the levels of invariance of the parameters of a test. The configural model is the simplest and least restrictive of all, since it only determines that, in the compared groups, the items of the test are explained by the same latent variables. The metric model comes next, since it determines that latent variables (configural model) and factor loadings are equal in the compared groups. The scalar model is more restricted than the other two, because it includes the previous restrictions and adds that the thresholds or intercepts of the items are equal (invariant) in the compared groups. Finally, the residual model contains all the previous restrictions and also determines that the residues are invariant in the compared groups. In this paper, as explained before, we will perform configural, metric, and scalar invariance analyses, but the residual invariance analysis will not be carried out.

Basically, the measurement invariance analysis consists of two steps. The first one evaluates the configural model fit indices. If the configural model is not rejected, then the second step can be carried out. In the latter, the fit of the more restrictive models is compared to the configural model fit, allowing us to infer whether or not there is invariance at the different levels represented by the more restrictive models.

In order to be considered invariant, the configural model must present at least one fit acceptable to the data. Usually, values of the Confirmatory Fit Index (CFI) ≥ 0.90 and Root Mean Square Error of Approximation (RMSEA) < 0.10 (Carmo, Brás, Batista, & Faísca, 2017) are deemed sufficient not to reject the model and, therefore, to conclude that the configural model is invariant (Schumacker & Lomax, 2016). In this paper, the aforementioned indices and cut-off points are used to evaluate the configural model fit.

While the first stage of invariance analysis is consensual and, therefore, relatively simple, the second stage of invariance analysis is relatively complex, because it is at this stage that the strong disagreements and controversies are found. One of the most widely used criteria to compare the more restrictive models to the configural model is the comparison of chi-squares (χ^2) and degrees of freedom, to identify statistically significant differences between the models. In this analysis, it is assumed that the more restrictive model is invariant when there are no statistically significant differences between the models. Usually, the cut-off point used to reject the more restrictive model has been $p < 0.05$. However, several studies indicate that the χ^2 differences and degrees of freedom criterion is very sensitive to the size of the groups analyzed, which may indicate statistically significant differences which, in fact, are negligible and should not be considered to reject the invariance of the more restricted model (Cheung & Rensvold, 2002; Putnick & Bornstein, 2016; Sass, Schmitt, & Marsh, 2014). Given the sensitivity of the χ^2 differences and degrees of freedom criterion, some authors proposed the use of other criteria. Among these criteria, probably the most widely used is the ΔCFI, that is, the comparison between CFI values (Putnick & Bornstein, 2016). Cheung and Rensvold (2002) state that $\Delta CFI \leq 0.01$ indicates that the more restrictive model should not be rejected, thus assuming the invariance of this model. However, the ΔCFI cut-off point is controversial. Meade, Johnson and Braddy (2008) suggest a much more rigorous cut-off point. For them, only $\Delta CFI \leq 0.002$ permits not to reject the more restrictive model. Chen (2007), in turn, suggests $\Delta CFI \leq 0.01$, combined with $\Delta RMSEA \leq 0.015$ and ΔSRMR (Standardized Root Mean Square Residual) ≤ 0.03 for metric invariance and $\Delta SRMR \leq 0.015$ for scalar or residual invariance.

Despite the importance of the fit indices to evaluate the invariance, there is no consensus so far on the best indices or the best cut-off points. It is the researcher who usually selects the criteria (Bowen & Masa, 2015). Considering the complexity pointed out, in this paper, both the χ^2 differences and degrees of freedom, and the ΔCFI criteria will be considered to compare the more restrictive models to the configural model. In this paper, the more restrictive model will be rejected if the χ^2 difference and degrees of freedom ($\Delta\chi^2$ (Δgl)) criterion shows $p < 0.01$ and the differences between the CFI criterion shows $\Delta CFI > 0.002$.

Another controversial methodological question in the measurement invariance analysis concerns what should be done when the more restrictive model does not seem to be invariant. The literature has suggested testing a partial invariance model that removes the constraint of the parameters that compromise the invariance of the rejected model. In this sense, the literature recommends investigating not only whether a test is invariant or not, but whether it presents partial invariance and whether this permits the comparison of its scores. It is always possible that partial invariance generates some kind of noise (Bowen & Masa, 2015). Because of that, Dimitrov (2010) recommends that less than 20% of the parameters be free. On the other hand, there are authors who suggest that at least half of the parameters should be invariant (Steenkamp & Baumgartner, 1998; Vandenberg & Lance, 2000). It is noteworthy that there are no empirical studies supporting these recommendations (Bowen & Masa, 2015). Thus, in this paper, whenever the metric or scalar invariance model is rejected, a partial invariance model that removes the constraints of the parameters that compromise invariance is created and tested.

All measurement invariance analyses in this study are performed using the R statistical software (version 3.5 R) (R Core Team, 2018) packages semTools (version 0.5-1) (Jorgensen, Pornprasertmanit, Schoemann, & Rosseel, 2018) and lavaan (version 0.6-3) (Rosseel, 2012). A confirmatory factor analysis of the items was carried out for the full sample to check the goodness of fit of the Metacognitive Monitoring Test factor structure. The model assumes a latent variable explaining the nine test items. Considering that the test items are dichotomous (0 and 1), the Weighted Least Square Mean and Variance Adjusted (WLSVM) estimator was used to perform the factor analysis of the items and the subsequent measurement invariance analyses. To verify the model's goodness of fit, we chose to reject the model when $CFI < 0.90$ and $RMSEA \geq 0.10$ (Carmo, et al., 2017). The model was considered to have good fit to the data when $CFI \geq 0.95$ and $RMSEA < 0.06$. In addition to the dimensionality analysis, Cronbach's alpha (1951) and Mcdonald's omega (1999) were calculated. After this analysis, the configural, metric and scalar invariance models were tested across educational level (higher education vs. 6th to 12th grades of basic education), sex (male vs. female) and nationality (Brazilian vs. Honduran) variables. The configural model was evaluated using the same rejection criteria used in the confirmatory factor analysis of the items for the full sample. If the configural model was not rejected, the analysis of invariance of the more restrictive models continued, comparing them to the configural model. The more restrictive model was not rejected if the values of the differences between the models presented $\Delta CFI \leq 0.002$ or p-value ≥ 0.01 for $\Delta\chi^2$ (Δdf)). The non-rejection of the model implied evidence of invariance at the level recommended by the model.

In case a more restrictive model was rejected, the constrained parameters that caused the model to be rejected were identified. These constrained parameters were relaxed so to generate a partial model with acceptable fit. The modus operandi was the following: Initially, we tried to verify if the removal of some constrained parameters of the more restrictive model allowed improving its goodness of fit. For this purpose, we used the lavaan package's lavTestScore function. This function allows you to perform the Lagrange Multiplier Test (Desjardins & Bulut, 2018), which consists of testing hypotheses to check if the model can improve its fit in case any constrained parameter is relaxed. *P* values < 0.05 indicate that the removal of some constrained parameter can improve the model's fit. The lavaan's lavTestScore and parTable functions allow verifying which relaxed parameter generates the greatest impact on the improvement of the model's fit, in terms of reduction of the model's χ^2. Only the constrained parameter with greater impact was removed, creating a new model, now with partial invariance. The procedure of comparing this model to the configural model was repeated. If the partial model was not rejected, the analysis was then completed. Otherwise, the constrained parameter that most impacted the rejection of the model was examined again. A new partial model was created, relaxing both the previous constrained parameter and the new constrained parameter that was identified, and so on, until the partial model was not rejected.

Results and Discussion

Prior to presenting the results of the confirmatory factor analysis of the items for the full sample and the measurement invariance analyses, it is important to present the distribution of the Metacognitive Monitoring Test raw scores, because it shows the wide performance diversity of the participants in this study. Moreover, this wide heterogeneity is verified across males and females, students of higher education and basic education, and Brazilian and Honduran students. This evidence is important, as it shows that the samples of this study represent an extensive diversity of possible performances in the test, thus encompassing different levels of metacognitive monitoring ability. As showed in Figure 1, the participants' performance reached all possible scores, ranging from 0 to 9 points.

The results of the confirmatory factor analysis of the items for the full sample and the measurement invariance analyses are presented in Table 1. The first result to be presented and discussed involves the confirmatory factor analysis of the items for the full sample, because this analysis tests the one-dimensional model that assumes the existence of a latent variable

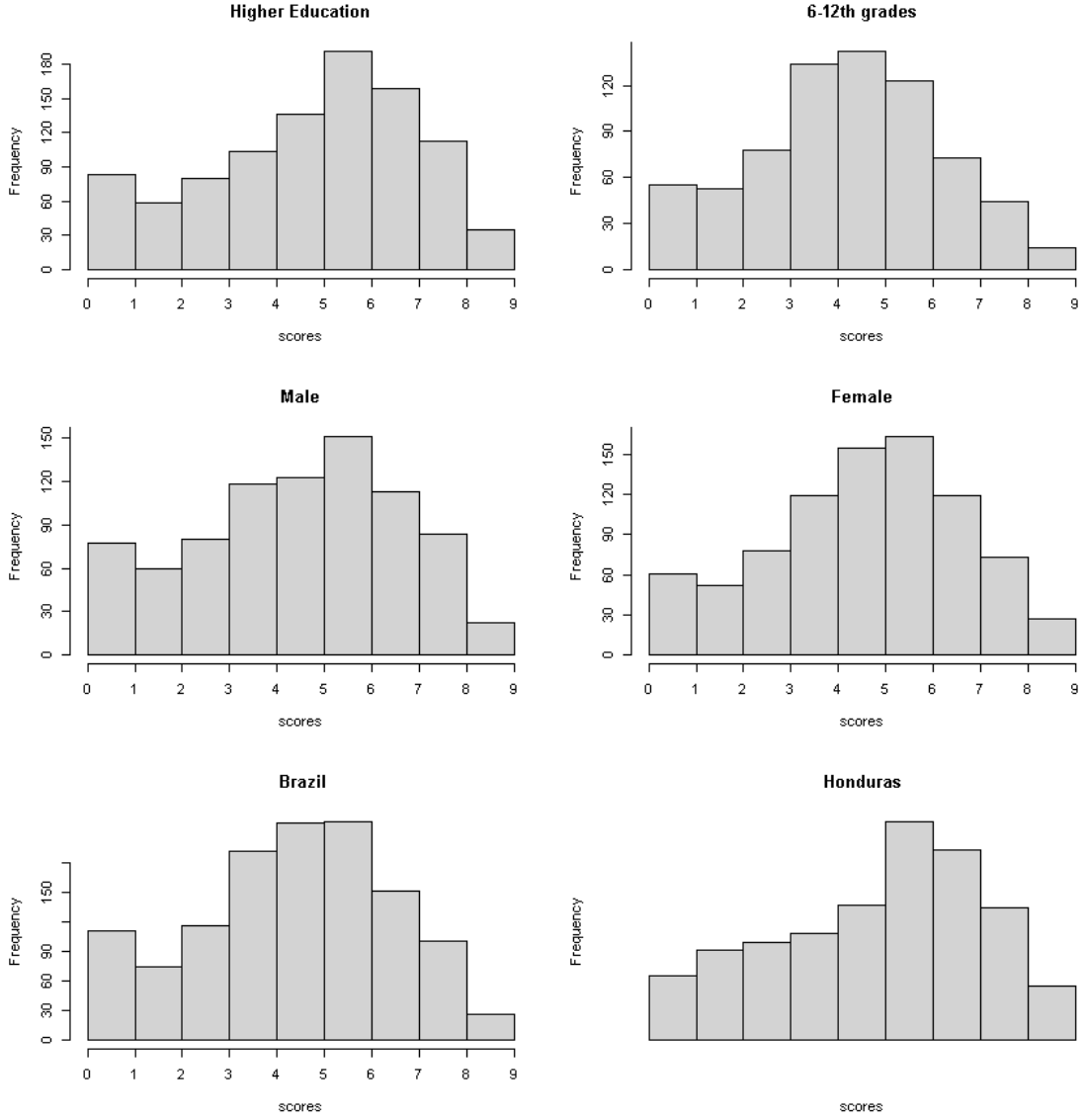

Figure 1. Histogram of the Metacognitive Monitoring Test raw scores distribution across sex, nationality, and educational level variables.

(metacognitive monitoring) to explain the variance of the nine test items. This analysis involving the full sample is of paramount importance, because if the one-dimensional model is rejected, it would not make sense to analyze the invariance of a model that had already been rejected.

The one-dimensional model presented good fit to the data for the full sample (see Table 1). In addition, the test items proved to be good markers of the metacognitive monitoring ability, as none of the factor loadings is less than 0.50, ranging from 0.50 to 0.82 (see Figure 2). The Metacognitive Monitoring Test reliability was 0.83 for Cronbach's alpha and 0.69 for Mcdonald's omega. The cut-off point traditionally used to evaluate the reliability of a test's scores is 0.70 (Viladrich, Angulo-Brunet, & Doval, 2017). Considering this cut-off point, the test score is acceptable according to the alpha result, and slightly below acceptable by the omega result. However, it is important to note that the omega index is much more demanding than Cronbach's alpha, since omega includes factor loadings in the reliability estimation, so that it is not unusual to find omega values considerably below the values indicated by the alpha. While there is already a relatively old and well-established consensus on the cut-off point of 0.70 for Cronbach's alpha values, the cut-off values for omega still lack further discussion and definition.

Contrasting the results of the analysis of the full sample of this study with those of previous studies, it should be noted that the favorable evidence for the one-dimensionality of the Metacognitive Monitoring Test in the full sample was already expected, because out of the three samples used in this study to make the full sample, one comes from Golino e Gomes (2011)'s study, and the other comes from Castillo (2018)'s study, and in both those studies there was indication

Table 1.
Results of the confirmatory factor analysis of the items and measurement invariance analysis across educational level, nationality and sex

Model	χ^2(df)	$\Delta\chi^2$(Δdf)	*p*	CFI	ΔCFI	RMSEA	IC 90%
Full sample	114(27)			**0.972**		**0.044**	**0.036-0.052**
Invariance across educational level							
Configural	166(54)			**0.967**		**0.050**	**0.041-0.058**
Metric	205(62)	39(8)	0.00	0.958	0.009	0.053	0.045-0.061
Partial Metric (item 8)*	186(61)	20(7)	**0.08**	0.963	0.005	0.049	0.041-0.058
Scalar	291(61)	125(7)	0.00	0.932	0.035	0.067	0.060-0.075
Partial Scalar (item 7)**	180(60)	14(6)	**0.04**	0.965	**0.002**	0.049	0.041-0.057
Invariance across nationality							
Configural	153(54)			**0.971**		**0.047**	**0.038-0.056**
Metric	204(62)	51(8)	0.00	0.959	0.012	0.052	0.045-0.060
Partial Metric (item 7)*	182(61)	29(7)	**0.02**	0.965	0.006	0.049	0.041-0.057
Scalar	348(61)	195(7)	0.00	0.917	0.054	0.075	0.067-0.083
Partial Scalar 1 (item 8)**	225(60)	72(6)	0.00	0.952	0.019	0.057	0.049-0.065
Partial Scalar 2 (items 3 and 8)**	179(59)	26(5)	0.00	0.965	0.006	0.049	0.041-0.058
Partial Scalar 3 (items 3, 7, and 8)**	169(58)	16(4)	**0.02**	0.968	0.003	0.048	0.040-0.056
Invariance across sex							
Configural	138(54)			**0.973**		**0.043**	**0.034-0.052**
Metric	145(62)	7(8)	**0.08**	0.974	**-0.001**	0.040	0.032-0.048
Scalar	153(61)	15(7)	**0.03**	0.971	**0.002**	0.042	0.034-0.051

Key: Bold terms indicate the non-rejected models and the values that support their non-rejection; χ^2= chi-square; df = degrees of freedom; Δ = difference; CI = confidence interval; * = factor loading removed, ** = threshold removed.

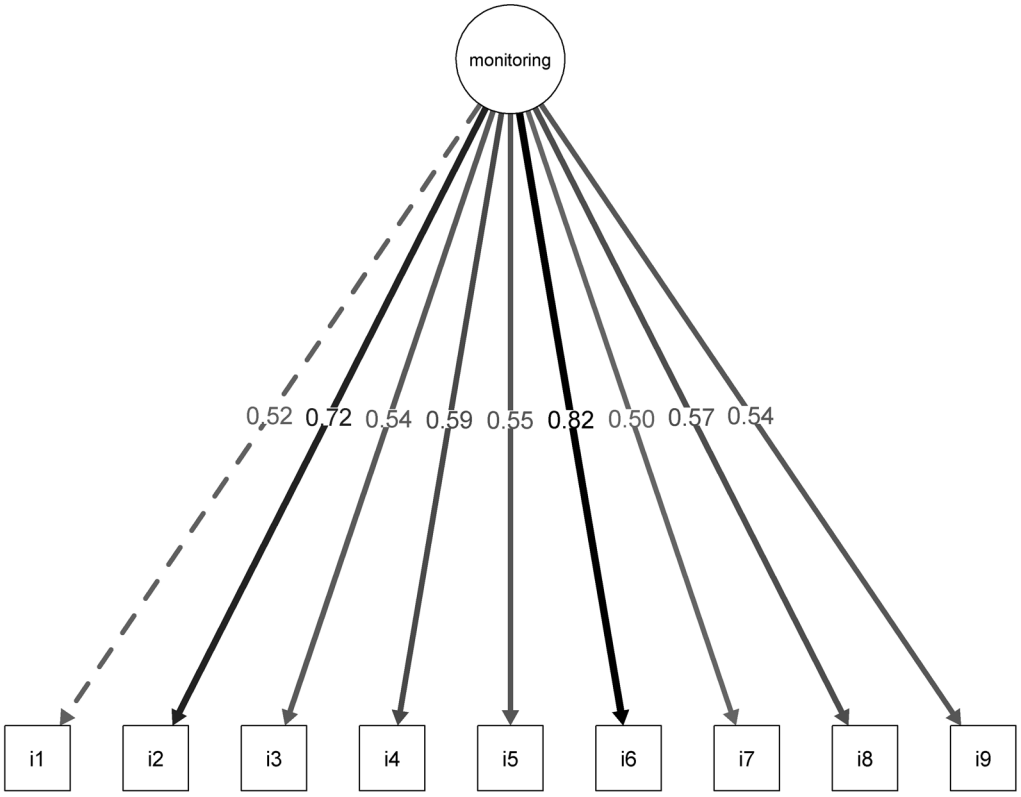

Figure 2. Structure of the Metacognitive Monitoring Test factor model.

of evidence favorable to the one-dimensionality of the test. Considering that the one-dimensional model was not rejected in the full sample, next are presented the results of the Metacognitive Monitoring Test invariance across the educational level, nationality, and sex variables, respectively.

The results indicate configural model invariance for the educational level variable, as this model presented good fit to the data (see CFI and RMSEA indices on Table 1). This indicates that only the monitoring latent variable explains the variance of the test items in the group of students from the 6th to the 12th grades of basic education and in the group of higher education students of this study's sample. The metric invariance model was rejected, but the partial metric invariance model that relaxed item 8's factorial loading constraint was not rejected (see bold values on Table 1), indicating that a small fit to the model provided the test with metric invariance. The scalar invariance model was rejected, but the partial scalar invariance model with item 7's thresholds relaxed was not rejected. As in the metric invariance model, a small fit to the scalar model enabled to achieve invariance.

Problem-posing, it is important to point out that it was expected that item 8's factorial loading parameter could compromise the scalar invariance, because this parameter proved to be problematic in the metric invariance, which is a more basic level than the scalar. However, this did not occur, indicating that a parameter can be problematic in a given model, without compromising the invariance of more restrictive levels. This result also shows that the methodological choice used in this article was wise by not starting the partial invariance of a more restrictive model by previously including the parameters removed from the previous restricted model. If this had been done, item 8's factorial load would have been mistakenly relaxed in the scalar invariance analysis.

To the extent that the partial scalar invariance model (item 7's threshold relaxed) is not rejected, this study shows evidence that it is possible to compare the performance of students from the 6th to the 12th grades of basic education in relation to higher education students, regarding the monitoring latent variable measured by the Metacognitive Monitoring Test. In so far as this study uses convenience samples, that is, not

representative of the populations of the groups analyzed, this study will not present any results regarding the difference between the groups.

The results also indicate configural model invariance for the nationality variable groups (see Table 1), implying that, for both Brazilians and Hondurans in the sample of this study, the monitoring latent variable alone explains the variance of the Metacognitive Monitoring Test items. The metric invariance model was rejected, but relaxing item 7's factor loading was enough to avoid rejection of the partial metric model. The scalar invariance model was also rejected. Three thresholds had to be relaxed (items 3, 7, and 8) to prevent the partial scalar model from being rejected. This non-rejected partial scalar invariance model brings evidence that the Metacognitive Monitoring Test allows comparing the performance of Hondurans and Brazilians, with regard to the monitoring latent variable.

Finally, the configural, metric and scalar invariance models for the sex variable groups were not rejected. This indicates favorable evidence that the scores of the Metacognitive Monitoring Test monitoring latent variable can be used to compare the performance of females and males.

Final Considerations

This study presents two major contributions to the field of metacognition studies. The first one involved investigating the Metacognitive Monitoring Test measurement invariance. As pointed out throughout the text, studies on metacognition have mostly used self-report questionnaires and think-aloud protocols, implying considerable damage to the quality of the measurement, since these methodologies generate substantial response and confirmation biases. The Metacognitive Monitoring Test is one of the few instruments in the field to measure metacognition based on the respondents' performance, and it was created intentionally with the purpose of generating a measurement without the response bias and confirmation bias. So far, the Metacognitive Monitoring Test had gone through the scrutiny of internal and external validity analyses, but the measurement invariance of this instrument had not been investigated.

The results found indicate evidence of configural, metric and scalar invariance for the sex variable, as well as configural invariance and partial metric invariance and scalar partial invariance for the nationality and educational level variables. The presence of scalar invariance across the three variables analyzed indicates favorable evidence that the Metacognitive Monitoring Test can be used to compare groups. Although the results indicate full scalar invariance only for the sex variable groups, and partial scalar invariance for the nationality and educational level variables groups, the partial scalar invariance involved set free just a few constrained parameters. For the educational level variable, only one parameter had to be relaxed, while for the nationality variable three parameters were relaxed. As previously explained, there is no consensus or reliable recommendation in the literature regarding the number of parameters removed for a partial invariance model to be considered adequate and not generate considerable bias. Nevertheless, the partial invariance models in this study meet Dimitrov (2010)'s strict criterion that a partial model is acceptable when it has no more than 20% of its parameters free. Other authors admit that partial models are acceptable as long as they have less than 50% of their parameters free (Steenkamp & Baumgartner, 1998; Vandenberg & Lance, 2000).

The second contribution of this study was presenting and discussing the major methodological challenges found in the measurement invariance analysis procedures. Measurement invariance analysis is a complex field and its stages demand considerable knowledge and methodological reflection from the researcher. This article brought to light a wide range of decision-making, and it also proposed some paths for invariance analysis that can be used by researchers in their investigations about instruments' invariance, such as the use of a 0.01 cut-off point instead of 0.05 for the p-value of the chi-square differences and degrees of freedom between the configural model and the more restrictive invariance models.

As for the study's limitations, the Metacognitive Monitoring Test invariance analyses performed herein used convenience samples, so the evidence found needs to be corroborated by future analyses, using diverse samples. This study focused on the measurement invariance analysis across three important variables, namely, nationality, educational level, and sex. However, the samples used did not necessarily show any balance in the variables focused on the analysis. For example, the Honduran sample includes only higher education students. In addition, considering that this study used Brazilians and Hondurans samples, it is important to examine whether evidence of invariance is also found across other nationalities. Finally, new studies may

incorporate other variables, for instance, the teaching setting variable (public vs. private), in order to bring broader evidence about the Metacognitive Monitoring Test measurement invariance.

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. doi:10.13189/ujp.2015.030104

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

Bártolo-Ribeiro, R., Simões, M. R., & Almeida, L. S. (2015). MetacognitiveAwarenessInventory (MAI): Adaptação e validação da versão portuguesa. *Revista Iberoamericana de Diagnóstico y Evaluación, 42*(2), 143–159. doi:10.21865

Bowen, N., & Masa, R. (2015). Conducting measurement invariance tests with ordinal data: A guide for social work researchers. *Journal of the Society for Social Work and Research, 6*(2). 2334–2315. doi:10.1086/681607

Carmo, C., Brás, M., Batista, L., & Faísca, L. (2017). Análise fatorial confirmatória da versão portuguesa da Escala Multidimensional de Perfecionismo de Frost. *Revista Iberoamericana de Diagnóstico y Evaluación*, *44*(2), 28–43. doi:10.21865/RIDEP44.2.03

Castillo, M. (2018). *Monitoramento e inteligência como preditores do desempenho acadêmico geral e específico no ensino superior*. Dissertação de mestrado, Universidade Federal de Minas Gerais, Belo Horizonte, MG, Brasil.

Chen, F. F. (2007). Sensitivity of goodness of fit indices to lack of measurement invariance. *Structural Equation Modeling*, *14*(3), 464–504. doi:10.1080/10705510701301834

Cheung, G. W., &Rensvold, R. B. (2002). Evaluating goodness of fit indices for testing measurement invariance. *Structural Equation Modeling*, *9*(2), 233–255. doi:10.1207/S15328007SEM0902_5

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

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika,16*(3), 297–334. doi:10.1007/BF02310555

Das-Smaal, E. A. (1990). Biases in categorization. *Advances in Psychology*, *68*, 349–387. doi:10.1016/S0166-4115(08)61332-1

Desjardins, C., &Bulut, O. (2018). *Handbook of educational measurement and psychometrics using R*. Boca Raton: CRC Press.

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

Dimitrov, D. M. (2010). Testing for factorial invariance in the context of construct validation. *Measurement and Evaluation in Counseling and Development*, *43*(2), 121–49. doi:10.1177/0748175610373459

Efklides, A., &Misailidi, P. (2010). *Trends and prospects in metacognition research*. Boston, MA: Springer US.

Fleming, S., & Frith, C. (2014). *The cognitive neuroscience of metacognition*. Berlin, Heidelberg: Springer Berlin Heidelberg.

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. doi: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. Retrieved from http://www.intestcom.org/files/ti26.pdf

Gomes, C. M. A., &Golino, H. F. (2014). Self-reports on students' learning processes are academic metacognitive knowledge. *Psicologia: Reflexão e Crítica*, *27*(3), 433–441. doi:10.1590/1678-7153.201427307

Gomes, C. M. A., Golino, H. F., & Menezes, I. (2014). Predicting school achievement rather than intelligence: Does metacognition matter? *Psychology*, *5*(9), 1095–1110. doi:10.4236/psych.2014.59122

Haukås, Å., Bjørke, C., &Dypedahl, M. (2018). *Metacognition in language learning and teaching*. New York, NY, London: Routlege.

International Test Commission. (2017). *The ITC guidelines for translating and adapting tests*, 2, 143–171.

Retrieved from http://www.intestcom.org/files/guideline_test_adaptation_2ed.pdf

Jorgensen, T. D., Pornprasertmanit, S., Schoemann, A. M., &Rosseel, Y. (2018). semTools: Useful tools for structural equation modeling. R package, version 0.5-1. Available from https://CRAN.R-project.org/package=semTools

Markman, E. M. (1977). Realizing that You Don't Understand: A Preliminary Investigation. Child Development, *48*(3), 986–992. doi:10.2307/1128350

Markman, E. M. (1979). Realizing that You Don't Understand: Elementary School Children's Awareness of Inconsistencies. Child Development, *50*(3), 643–655. doi:10.2307/1128929

McDonald, R. P. (1999). *Test theory: A unified treatment*. Mahwah, NJ: Erlbaum.

Meade, A. W., Johnson, E. C., &Braddy, P. W. (2008). Power and sensitivity of alternative fit indices in tests of measurement invariance. *Journal of Applied Psychology*, *93*(3), 568–592. doi:10.1037/0021-9010.93.3.568

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. doi:10.1007/s10212-010-0040-7

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. doi:10.1007/s11409-018-9183-8

Peña-Ayala, A. (2015) *Metacognition: Fundaments, Applications, and Trends. A Profile of the Current State-Of-The-Art*. Springer New York.

Pintrich, P. R., Wolters, C. A., & Baxter, G. P. (2000). Assessing metacognition and self-regulated learning. In G. Schraw & J. C. Impara (Eds.), Assessing metacognition and self-regulated learning (pp. 43–97). Lincoln, NE: Buros Institute of Mental Measurements.

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. doi:10.1016/j.dr.2016.06.004

R Core Team. (2018). R (version 3.5). [Computer software]. Vienna, Austria: R Foundation for Statistical Computing. Available from http://www.r-project.org/

Rosseel, Y. (2012). lavaan: An R Package for Structural Equation Modeling. R package, version 0.6.3. Available from http://www.jstatsoft.org/v48/i02/

Sass, D., Schmitt, T., & Marsh, H. (2014). Evaluating model fit with ordered categorical data within a measurement invariance framework: A comparison of estimators. *Structural Equation Modeling:A Multidisciplinary Journal*, *21*(2), 167–180, doi:10.1080/10705511.2014

Schumacker, R., & Lomax, R. (2016). A beginner's guide to *structural equation modeling. New York: Routledge.*

Steenkamp, J-B., & Baumgartner, H. (1998). Assessing measurement invariance in cross-national consumer research. *Journal of Consumer Research*, *25*(1), 78–90. doi:10.1086/209528

Vandenberg, R. J., & Lance, C. E. (2000). A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. *Organizational Research Methods*, *3*, 4–69. doi:10.1177/109442810031002

Viladrich, C., Angulo-Brunet, A., &Doval, E. (2017). A journey around alpha and omega to estimate internal consistency reliability. *Anales de Psicología, 33*(3),755–782. doi:10.6018/analesps.33.3.268401

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). New York, NY: Oxford University Press.

Recebido em: 05/08/2019
Reformulado em: 14/09/2020
Aprovado em: 09/11/2020

About the authors:

Cristiano Mauro Assis Gomes is a Psychologist. Doctor in Education and Post-PhD in Educational Psychology. Professor of the Psychology Department, the Graduate Program in Psychology: Cognition and Behavior, and the Graduate Program in Neuroscience at Universidade Federal de Minas Gerais (UFMG), Brazil. Head of the Cognitive Architecture Research Laboratory (LAICO). CNPq productivity fellow.
ORCID: https://orcid.org/0000-0003-3939-5807
E-mail: cristianomaurogomes@gmail.com

Jhonys de Araujo is a Biologist. Master in Neuroscience and doctoral student in Psychology: Cognition and Behavior at Universidade Federal de Minas Gerais (UFMG), Brazil. Researcher of the Cognitive Architecture Research Laboratory (LAICO).
ORCID: https://orcid.org/0000-0002-7936-7440
E-mail: jhonys.bio@gmail.com

Marcio Alexander Castillo-Díaz is a Psychologist. Master in Developmental Psychology and Doctor in Psychology: Cognition and Behavior at Universidade Federal de Minas Gerais (UFMG), Brazil. Researcher of the Cognitive Architecture Research Laboratory (LAICO). Professor of the Student Affairs Department and the Graduate Program in Psychometrics and Educational Evaluation at Universidad Nacional Autónoma de Honduras (UNAH), Honduras.
ORCID: https://orcid.org/0000-0002-9489-7036
E-mail: marcio.castillo@unah.edu.hn

Contact:

Laboratório de Investigação da Arquitetura Cognitiva (LaiCo)/ FAFICH-UFMG. Campus Pampulha
Av. Pres. Antônio Carlos, 6627, sala 4036, Pampulha
Belo Horizonte-MG, Brazil
CEP: 31270-901