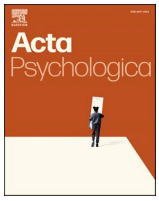
Acta Psychologica 246 (2024) 104291

Contents lists available at ScienceDirect

Acta Psychologica

journal homepage: www.elsevier.com/locate/actpsy



Measurement invariance across countries of the Test of Memory Strategies (TMS): A contribution to the cross-national validity study

Roberto Giorgini [a], Fernando Maestu [b,c], Fernandes Margarida Sara [d], Massimiliano Pastore [e], Maria Abellan [c], Andrea Quattrone [f], Sara Caparello [g], Aldo Quattrone [h], Maria Grazia Vaccaro [f,h,*]

[a] Department of Experimental and Clinical Medicine, University Magna Grœcia of Catanzaro, Italy
[b] Networking Research Center on Bioengineering, Biomaterials and Nanomedicine (CIBER-BBN), Complutense University of Madrid, Spain
[c] Department of Experimental Psychology, Faculty of Psychology, Complutense University of Madrid, Spain
[d] Department of Psychology and Education, CINTESIS – Research Center For Technology and Health Services– Portucalense University, Portugal
[e] Department of Developmental and Social Psychology, University of Padua, Italy
[f] Department of Medical and Surgical Sciences, Magna Graecia University of Catanzaro, Italy
[g] Department of Business and Legal Sciences, University of Calabria, Italy
[h] Neuroscience Research Center (CR), Department of Medical and Surgical Scienze, Magna Graecia University of Catanzaro, Italy

ARTICLE INFO

Keywords:
Measurement invariance
Test of Memory Strategies
Memory
Executive functions
Construct validity
Structural equation modeling

ABSTRACT

Previous literature showed a complex interpretation of recall tasks due to the complex relationship between Executive Functions (EF) and Long Term Memory (M). The Test of Memory Strategies (TMS) could be useful for assessing this issue, because it evaluates EF and M simultaneously. This study aims to explore the validity of the TMS structure, comparing the models proposed by Vaccaro et al. (2022) and evaluating the measurement invariance according to three countries (Italy, Spain, and Portugal) through Confirmatory Factor Analysis (CFA). Four hundred thirty-one healthy subjects (Age mean = 54.84, sd = 20.43; Education mean = 8.85, sd =4.05; M = 177, F = 259) were recruited in three countries (Italy, Spain, and Portugal).

Measurement invariance across three country groups was evaluated through Structural Equation modeling. Also, convergent and divergent validity were examined through the correlation between TMS and classical neuropsychological tests. CFA outcomes suggested that the best model was the *three-dimensional model,* in which list 1 and list2 reflect EF, list 3 reflects a mixed factor of EF and M (EFM) and list4 and list5 reflect M. This result is in line with the theory that TMS decreases EF components progressively. TMS was metric invariant to the country, but scalar invariance was not tenable. Finally, the *factor scores* of TMS showed convergent validity with the classical neuropsychological tests.

The overall results support cross-validation of TMS in the three countries considered.

1. Introduction

Deficits in Executive Functions (EF) and Long-Term Memory (M) are common in different neurological and psychiatric diseases, but their contribution to test performance is difficult to single out (Higginson et al., 2003; Chang et al., 2010; Fossati et al., 1999; Duff et al., 2005; Busch et al., 2005; Craik et al., 2018). There are many definitions and operationalizations of EF, but many tests used as an observed variable (OV; like in the Structural Equations Models framework) present the same phenomenon: *task impurity problem* (Baggetta & Alexander, 2016; Miyake & Friedman, 2012). When a single task assesses different EF components or other cognitive functions, the interpretation of task measure is difficult to make, this situation is referred to as *task impurity problem.* Although there are pieces of evidence of double dissociation between EF and M (Glisky et al., 1995; Postle et al., 1999), the evaluation of recovery problems through neuropsychological tests is still hazy, indeed different processes of the same cognitive function (e.g. Working Memory) are mediated by non-mnemonic executive control and mnemonic processes (Baddeley, 1992; Postle et al., 1999). For example, in Digit Span Backward (Wechsler, 1981), the subject must manipulate the

* Corresponding author at: Neuroscience Research Center (CR), Department of Medical and Surgical Science, Magna Graecia University, Catanzaro, Italy.
E-mail address: mg.vaccaro@unicz.it (M.G. Vaccaro).

https://doi.org/10.1016/j.actpsy.2024.104291
Received 13 February 2024; Received in revised form 28 March 2024; Accepted 22 April 2024

information (Executive process of *Working Memory*, EF) while maintaining the data in Short Term Memory storage. In another impure task, like the Verbal Fluency Task, the subject must recover the information (EF) in Long Term Memory storage (M) and change the word generator criterion (first letter) after one minute (*Shifting*, EF; Miyake et al., 2000).

In the majority of tasks that investigate EF and M, there is not an unmixed effect of these two *factors,* thus it is hard to determine a causal effect between them for clinical applications (Craik et al., 2018). Given this, accurate neuropsychological tests are required to define if a pathological memory condition is related to EF and/or M.

Chang et al. (2010), demonstrated that Mild Cognitive Impairment (MCI) patients with high scores in EF showed better performance in verbal memory tasks than MCI patients with lower scores in EF. This was also associated with cortex volume differences between groups. The outcomes of instruments used to estimate EF and M could be a predictor of dementia comorbidity in different neurological diseases and a prognostic indicator in neurodegenerative pathologies. For example, Levy et al. (2002) analyzed in a longitudinal study the evolution of dementia related to neuropsychological tests in idiopathic Parkinson's disease patients, their results showed that the principal predictors were the level of EF and M at baseline.

Thus, it is crucial to develop neuropsychological tests that detect primary memory deficit or impairment of EF when a subject shows a failure in a recall task.

The Test of Memory Strategies (TMS), an immediate recall task, was developed by Yubero et al. (2011) using a different methodology. Specifically, TMS gradually reduces the EF processes in verbal learning, in contrast to traditional neuropsychological tests that assess EF and M independently. The design of TMS includes fifty words organized in five lists, which progressively reduce the EF components, from list-1 to list-5.

The TMS showed adequate psychometric properties (Fernandes et al., 2018; Vaccaro et al., 2022; Yubero et al., 2011) the correlation between TMS and classical neuropsychological tests supported convergent validity and the high internal consistency promoted the reliability.

Furthermore, Fernandes et al. (2018), performed a *Principal Components Analysis* (PCA) to assess the construct validity of TMS. The results suggested a *two-factor* structure, but the loading of list 4 was complex. To assess the *complex structure* issue, Vaccaro et al. (2022) conducted a *Confirmatory Factor Analysis* (CFA) where the loadings were constrained for specific lists*;* their results supported the hypothesized *two-factor* structure through a model comparison. Moreover, TMS was used in pilot clinical studies in which the results showed a significant difference between mixed clinical samples and healthy subjects, considering a possible clinical application of TMS (García-Laredo et al., 2021; Yubero et al., 2011).

The next step in the *Classical Test Theory* framework is to create normative values, deleting the effect of demographic variables detected in the TMS scores (Vaccaro et al., 2022; Yubero et al., 2011), but first, the *measurement invariance* across countries of TMS should be tenable. Although the original TMS was developed in Spain and later translated into Italian and Portuguese, it has never been verified if the test retains varying degrees of invariance throughout nations.

A lack of *measurement invariance* on different levels could induce an interpretation *bias* of measure (Gregorich, 2006), so before TMS will be applied in clinical paradigms to build normative values country-related, the test should have the same structure across groups (*configural invariance*) and the same factor loadings magnitude across groups *(metric invariance*). Furthermore, to compare the country groups, the *scalar invariance* should be defendable (same intercepts across groups), because different intercepts could suggest some advantage of one group due to other variables not being considered.

The study aims to explore the construct validity of TMS, comparing the models proposed by Vaccaro et al. (2022) and evaluating the *measurement invariance* across three country groups (Italy, Spain, and Portugal) to support future clinical validation of TMS.

2. Materials and method

All statistical analyses were performed in the statistical programming environment R (R core Team, 2021). The CFA and Structural Equation Models were performed through *lavaan* package (Rosseel, 2012).

2.1. Participants

The TMS was individually administered to 436 healthy subjects in three countries: Italy ($N = 121$), Portugal ($N = 135$), and Spain ($N = 180$). The inclusion criteria required that participants had to be in good health and not have any cognitive impairment that would interfere with their daily life activities. The Mini-Mental State Examination (MMSE) was administered to screen for the presence of cognitive impairment, using the validated cutoffs for each country.

Italian participants were recruited among patients familiar and staff of the Movement Disorder Unit of the Local University Hospital; Spanish participants were enrolled from a local private organization and research institute; Portuguese participants were recruited from senior universities and daycare centers and a sample of younger participants was recruited using snowball sampling.

2.2. Assessment

We selected all tests country-adapted in common with the three data sets.

MMSE was included for general cognitive functioning (Folstein et al., 1975; Lobo et al., 1979; Magni et al., 1996; Morgado et al., 2009). Verbal Phonological Fluency was chosen to evaluate *Verbal Memory Access* and *Shifting* abilities. (COWAT; Benton & Hamsher, 1978; Wechsler et al., 2012; Cavaco et al., 2013; Bianchi & Dai Prà, 2008) and Semantic Verbal Fluency (SVF; Wechsler et al., 2012; Cavaco et al., 2013; Bianchi & Dai Prà, 2008). Digit span Forward and Backward (Wechsler, 1981; Wechsler et al., 2012; Rocha et al., 2008; Orsini & Pezzuti, 2013; Orsini & Pezzuti, 2015) were selected as a mesasure of short term memory span. Lastly, executive functions—specifically, inhibition—were evaluated using the Stroop test (Stroop, 1935; Wechsler et al., 2012; Fernandes, 2013; Orsini & Pezzuti, 2013; Orsini & Pezzuti, 2015; Scarpina & Tagini, 2017).

2.2.1. TMS

Five-word lists are provided by TMS, and following each presentation, participants are instructed to repeat as many words as they can. Ten words from List 1 lack phonetic and semantic links. The structure of lists 1 and 2 is the same.

List 3 has the words arranged into two semantic categories; however the words are presented in a random sequence unrelated to the categories (furniture and trees).

In contrast, list 4 displays the terms in a hierarchy related to two categories (labor tools and transportation). Lastly, list 5 shares list 4's structure, but participants are informed of the existence of two semantic categories (sports and vegetables) in distinct instructions.

2.3. Statistical analysis

To evaluate the construct validity of TMS, we conducted a Confirmatory Factor Analysis (CFA) using a maximum likelihood robust (MLR). We compared the five models proposed by Vaccaro et al. (2022), which include: the *unidimensional model* (M1); a *two-factor model* in which EF were reflected by list1 and list2, while M by list3, list4, and list5 (M2)*; three-dimensional model,* a modified *two-factor structure* where list 3 loadings on mixed Executive Functions-Memory (EFM; M3); *alternative two-factor model* in which list 3 reflects EF (M4); a second *alternative two-factor model* where all the variables load onto EF, except list 5 (M5). All models are shown in Fig. 1.

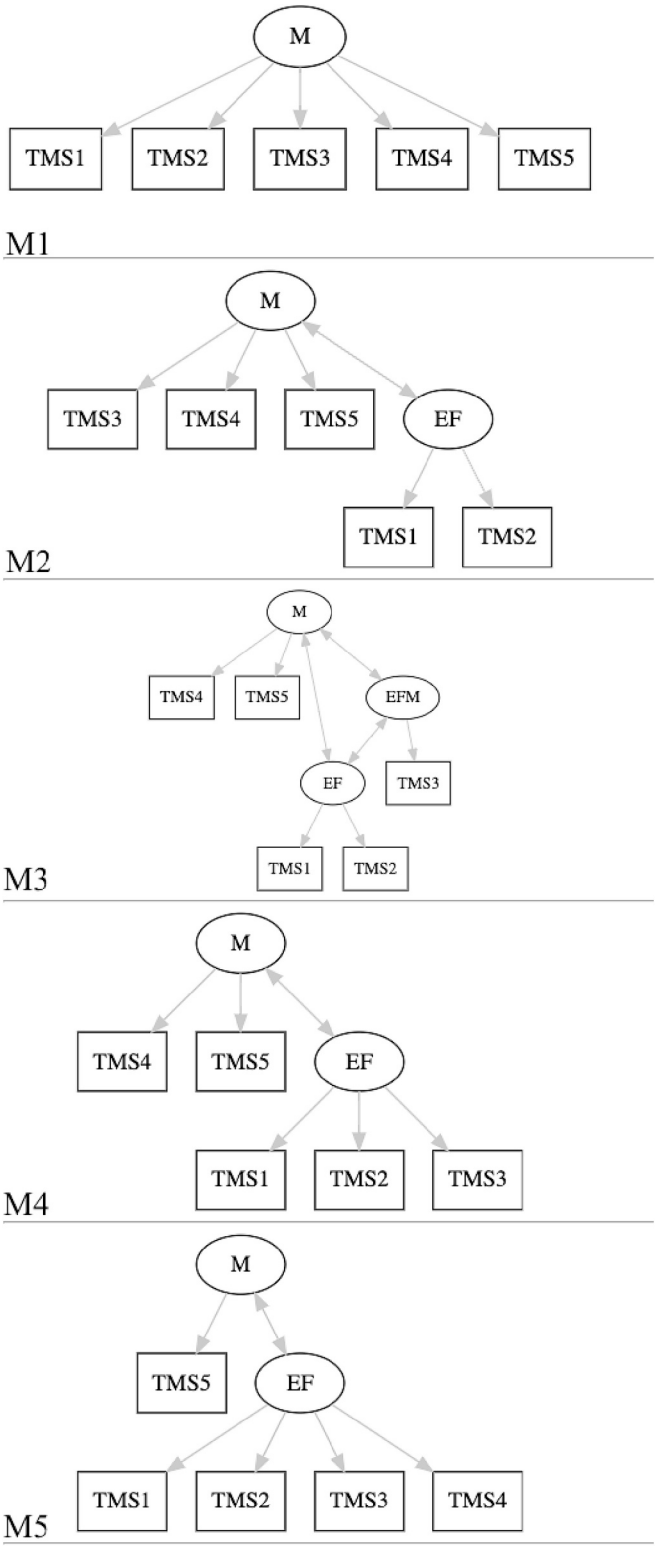


Fig. 1. Path diagram of the models.
Note. M1 = *unidimensional model*; M2 = *two-factor model;* M3 = *three-dimensional model*; M4 = *alternative two-factor model*; M5 = second *alternative two-factor model;* TMS1 = total score of list 1 of TMS*;* TMS2 = total score of list 2 of TMS*;* TMS3 = total score of list 3 of TMS*;* TMS4 = total score of list 4 of TMS*;* TMS5 = total score of list 5 of TMS.

We considered different fit indices for all models: chi-squared ($\chi 2$), *Robust Comparative fit index* (CFI robust), Robust *Tucker Lewis index* (TLI robust), Robust *Root Mean Squared Error of Approximation* (RMSEA robust), the *Standardized Root Mean Squared Residual* (SRMR) and the Akaike Weights *w*(AIC). The *w*(AIC) is the normalized relative likelihood of M_i, it may be seen as the likelihood that, among a particular collection of models calculated using the same data, M_i is the best model (Wagenmakers & Farrell, 2004). Therefore, in the set of considered models, the best will be the one with highest *w*(AIC) (Wagenmakers & Farrell, 2004).

The goodness of fit was considered if CFI and TLI were $>$ 0.95, RMSEA $<$0.06, and SRMR $<$0.08 (Hu & Bentler, 1999).

After identifying the best TMS model, we used it to evaluate its invariance across countries.

First, we assessed configural invariance (CINV), meaning that only the model structure was constrained to be equal across groups. As a second step, we assessed metric invariance (MINV) by constraining the factor loadings to be equal across groups. Finally, we assessed the scalar invariance (SINV) by constraining the factor loadings and intercepts to be equal across the groups. Configural and metric invariance ensure that items are equally weighted across countries when computing the subscores.

Additionally, scalar invariance should be examined when using scales to assess mean differences between groups since, if it is absent, the difference in group means may not accurately reflect variations in LV (Gregorich, 2006).

Measurement invariance across countries was assessed by considering the differences in CFI and RMSEA indices obtained with the different types of invariances. A change ≤ -0.005 between CFI (ΔCFI) and a change $\geq$0.015 in RMSEA (ΔRMSEA) were considered indicators of *metric* non-invariance, while ΔCFI ≤ -0.005 and ΔRMSEA $\geq$0.010 were considered indicators of *scalar* non-invariance (Chen, 2007).

Correlations between TMS *factor scores* and traditional outcomes of neuropsychological tests were conducted to evaluate convergent and divergent validity and ensure high precision using weighted ratings. Finally, the correlation between TMS *factor scores* and Phonological fluency was calculated as an index of nomological validity (Whiteside et al., 2016).

3. Results

Descriptive statistics for the specific sample, including demographic characteristics such as gender, age, and education (in years) are shown in Table 1.

3.1. CFA results

The outcomes of CFA suggest that the best models were M3, M2, and M4, despite all models having excellent fit indices (see Table 2).

The best model was M3, with a weight of 0.45 and relative evidence approximately 1.88 times that of models M2 and M4. Therefore, we chose M3 as the best model. Table 3 displays M3 factor loadings. The estimated covariance between EF and EFM was 0.678, between EFM and M 0.685, and between EF and M 0.865.

Table 1
Descriptive statistics of demographic variables.

	Sex		Age	Education
Italy	M = 47	F = 74	45.90 ± 20.36	13.17 ± 3.97
Spain	M = 79	F = 101	53.99 ± 16.66	7.37 ± 2.41
Portugal	M = 51	F = 84	63.99 ± 21.35	8.14 ± 4.54

Note 1 M = male; F = female.

Table 2
Models comparison.

	CFI. robst	TLI. robust	RMSEA. robust	SRMR	*w* (AIC)	*χ2*	df
M1	0.991	0.983	0.052	0.022	0.03	10.961 $p = .05$	5
M2	0.999	0.997	0.023	0.014	0.24	4.843 $p = .30$	4
M3	1.000	1.007	0.000	0.008	0.45	1.462 $p = .69$	3
M4	0.999	0.997	0.023	0.014	0.24	4.855 p = .30	4
M5	0.991	0.983	0.052	0.022	0.03	10.961 p = .05	5

M1 = unidimensional model; M2 = two-factor model; M3 = three-dimensional model; M4 = alternative two factor model; M5 = second alternative two factor model; CFI.robust = Robust Comparative fit index; TLI.robust = Robust Tucker Lewis index;, RMSEA.robust = Robust Root Mean Squared Error of Approximation; SRMR = Standardized Root Mean Squared Residual; w(AIC) = Akaike weights;χ2 = chi-squared; df = degrees of freedom.

Table 3
M3 factor loading.

Latent variable	List	Estimate	SE	Stand.estimate	*p*
EF	TMS1	0.89	0.072	0.649	$p < .001$
	TMS2	0.98	0.064	0.730	$p < .001$
EFM	TMS3	1.51	0.056	1.000	$p < .001$
M	TMS4	1.36	0.080	0.820	$p < .001$
	TMS5	1.32	0.089	0.684	$p < .001$

EF = Executive Functions; EFM = Executive-Function-Memory;M = Memory; TMS1 = sum score of total word remembered from list 1;TMS2 = sum score of total word remembered from list 2;TMS3 = sum score of total word remembered from list 3;TMS4 = sum score of total word remembered from list 4; TMS5 = sum score of total word remembered from list 5.

3.2. Measurement invariance across countries results

In Table 4, the fit indices of invariance models are displayed.

For CINV, the estimated fit indices were: $\chi2 = 13.845$ (df = 9, $p = .128$), CFI robust = 0.0.991, TLI robust = 0.0.969, RMSA robust = 0.065 and SRMR = 0.022.

For MINV, the estimated fit indices were: $\chi2 = 16.515$ (df = 13, $p = .222$), CFI robust=. 0.993, TLI robust = 0.984, RMSA robust = 0.046 and SRMR = 0.031.

Finally, for SINV, the estimated fit indices were: $\chi2 = 76.089$ (df = 17, $p < .001$), CFI robust=. 0.898, TLI robust = 0.820, RMSA robust = 0.155 and SRMR = 0.085.

Comparison between CINV and MINV suggested that MINV did not fit worse compared to CINV, so the metric invariance between country groups was tenable ($\chi2$ *diff* = 2.399, $p = .662$). Also, ΔCFI and ΔRMSEA supported the results of the comparison based on $\chi2$ (see Table 4). On the other hand, the comparison between MINV and SINV suggested that scalar invariance across the three countries considered was not tenable ($\chi2$ *diff* = 60.458, $p < .001$), this result is reinforced by ΔCFI = 0.095 and ΔRMSEA = 0.109.

3.3. Convergent and divergent validity of TMS results

Fig. 2 displays the correlations between the TMS *factor scores* and the raw scores from standard neuropsychological tests. We found positive correlations between EF *factor scores*, Digit Span forward/backward, Stroop interference score, COWAT, and SVF. M *factor scores* showed positive correlations with Digit Span forward/backward, COWAT, and Stroop interference scores. Finally, EFM *factor scores* exhibited positive correlations with Digit Span forward/backward, Stroop interference score, and COWAT. None of the three-factor *scores* was related to MMSE, the magnitude of correlation between TMS sub-scores and MMSE was very small ($0.16 < r < 0.18$).

4. Discussion

The degree of importance for producing results from the traditional neuropsychological tests between Executive Functions (EF) and Long-Term Memory (M) is still uncertain. This is probably due to the *task impurity problem,* namely the usage of tests in which outcomes are produced by the mixed measure of EF and/or other cognitive functions (Baggetta & Alexander, 2016; Miyake & Friedman, 2012). Researchers might create tests that are more discriminating or offer scales that are based on different methods to avoid the task impurity problem.

Test of Memory Strategies (TMS; Yubero et al., 2011) is based on a different viewpoint than classical neuropsychological tests, indeed TMS proposes to assess EF and M together, decreasing EF components through the recall of words semantically organized.

This study aimed to evaluate the construct validity of TMS through Confirmatory Factor Analysis (CFA) and to analyze the *measurement invariance* across three country groups (Italy, Portugal, and Spain).

Our results suggested that the best fitting model between the models proposed by Vaccaro et al. (2022) was the *three-dimensional model* (M3), where list 3 of TMS reflects a mixed measure between EF and M (EFM), lists 1 &2 reflect EF and list 4&5 reflect M.

This outcome is partially in line with previous literature about TMS. The first possible explanation could be related to the different sample sizes between studies of Vaccaro et al. (2022) and Fernandes et al. (2018). Although the sample size in the current study was larger than in previous research, this does not necessarily imply that it is superior but could explain the discrepancy noted in the model comparison findings.

The little variation from Vaccaro et al. (2022) results is most likely due to sample size; they conducted the same analysis as this study, and there isn't much of a difference between the two model comparisons.

Fernandes et al. (2018) performed Principal *Components Analysis* (PCA) on TMS, in this statistical approach the Latent Variables (LV) are predicted by Observed Variables (OV) multiplied by regression coefficients, also the standard error is not included in the model. Their findings showed a two-factor model with a complex loading on List 4, this is not in line with the results of the present study, because in the best-fitting model, List 4 is constrained to load only onto M with no cross-loading. A possible explanation for the difference in our results from Fernandes et al. (2018) is the different conceptualization and, accordingly, dissimilar computation of the model.

In our study, the structure of TMS is intended as a reflective model, thus the causal direction of the path is inverted than in PCA. In other words, the three factors found are underlying dimensions that cause the OV (and vice versa in PCA).

Theoretically, the TMS progressively decreases EF components from List 1 to 5, so the finding of List 3 loads onto a theoretic mixed factor (EFM) increases the construct validity of the test and supports the notion that measures of immediate recall tests reflect EF and M together. EFM

Table 4
Fit indices for invariance testing ($n = 431$).

Model	*χ2*	df	Δ*χ2*	Δdf	*p*	CFI	ΔCFI	RMSEA	ΔRMSEA
CINV	13.845	9				0.991		0.065	
MINV	16.515	13	2.60	4	0.625	0.993	−0.002	0.046	−0.019
SINV	76.089	17	60.45	4	<0.001	0.898	0.095	0.155	0.109

CINV = Configural invariance model; MINV = metric invariance model; SINV = scalar invariance model; *χ2* = chi-squared; df = degrees of freedom; CFI = *Robust Comparative fit index*; TLI = Robust *Tucker Lewis index;* RMSEA = Robust *Root Mean Squared Error of Approximation*.

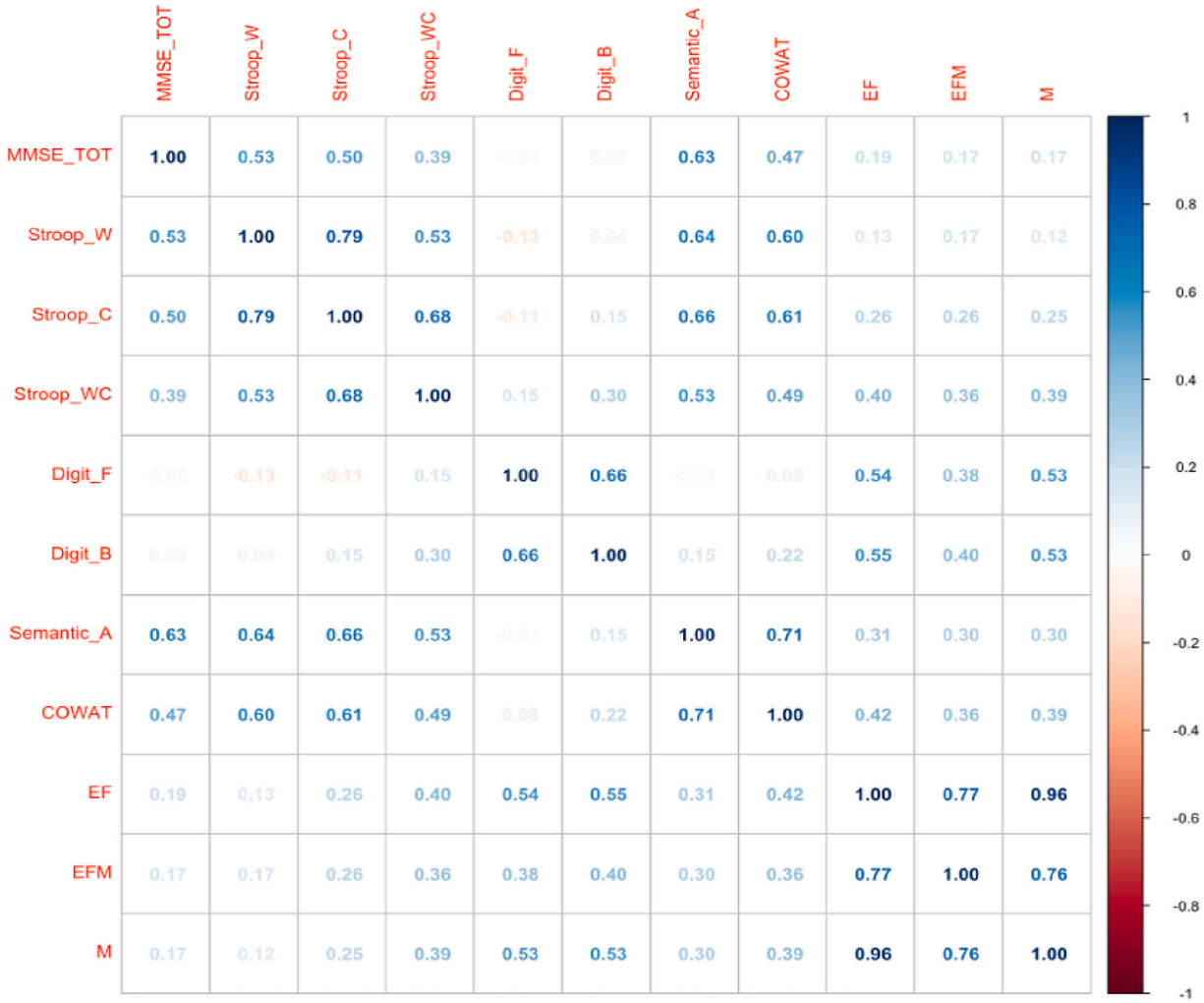


Fig. 2. Correlations between TMS factor score and classical neuropsychological tests raw scores (n = 431).
Note. MMSE_TOT = Mini-Mental State Examination raw total score; Stroop_W = raw total score of reading word task of Stroop test; Stroop_C = raw total score of reading color task of Stroop test; Stroop_WC = raw total score of interference task of Stroop; Digit_F = Digit Span forward; Digit_B = Digit Span backward; Semantic_A = Semantic Fluency task; COWAT: Verbal Fluency task; EF = TMS Executive Functions *factor score*; EFM = Executive Functions and Memory *factor score; M* = Memory *factor score*.

was interpreted as a mixed factor because, given that the elements in List 3 are arranged by an implicit cue (which reduces the involvement of EF in the coding strategy), the items are presented at random, and as a result, participants must develop self-strategies to identify this cue (Yubero et al., 2011). Empirically, the only support of this hypothesis was the positive covariance between EFM and the other factors that emerged from M3, at least they might be viewed as interdependent entities but not the same things, because the covariance was moderate. This is not enough to strongly argue that these three constructs are distinct, future studies are needed to assess this issue. The covariance between EF and M could lead to the idea that these latent variables are the same (Rönkkö & Cho, 2022). However, the value of this covariance doesn't reach the cut-off suggested by Rönkkö and Cho (2022).

To extend TMS's validity, we evaluated the *measurement invariance* according to countries, through Structural Equation modeling. The results suggest that TMS configural invariance was tenable, further TMS was metric invariant to country groups, a crucial prerequisite for relevant cross-cultural validation (Gregorich, 2006). Instead, TMS is not scalar invariant, namely, some groups do not have the same intercepts in the model. A lack of *scalar invariance* could produce interpretation bias in group comparisons due to the presence of confounding variables, indeed some groups seem advantaged to others at the same levels of the LV. In detail, Portuguese participants display higher intercepts than Italians and Spanish in each OV of TMS. A possible interpretation concerns a difference in age and education between groups, but this is unlikely, as the Portuguese sample showed higher levels of age, which is negatively related to EF and M performances. Analogously, compared to the Italian and Spanish participants, the Portuguese sample showed lower mean educational levels, which is positively related to EF and M performances.

A probable explanation of these results is a word's different frequency of occurrence across countries. We found a discrete number of phonological errors or intrusions in the Italian sample, this phenomenon encourages that some words have a low frequency of occurrence. Indeed, in a given language, people are more likely to remember a phonetically similar word with a higher frequency of occurrence when the correct word has a low frequency of occurrence (Roodenrys et al., 2002).

According to the literature, TMS' EF factor scores have the greatest correlation coefficients with Digit Span (forward and backward), interference Stroop measure, and Verbal Fluency Task. Consistently, also TMS' M factor scores have a discrete correlation with Digit Span and verbal fluency tasks; we discuss M and EFM convergent validity in limitation.

In previous studies, TMS showed good validity and reliability (Fernandes et al., 2018; Vaccaro et al., 2022), we found that the relationships between LV and OV (loadings) are equal between country groups, supporting the overall validity of TMS. The presence of *metric invariance* of the scale supports the cross-cultural validation in the three countries considered (Italy, Portugal, and Spain) because theoretically, the LV have the same meaning across the groups (Gregorich, 2006), and the items have similar weights across groups, according to empirical data. The main TMS application regards the clinic practice, since the first administration the outcomes of TMS have shown accordance with the specific deficit of the clinical sub-samples involved. Indeed, patients with frontal impairment showed lower scores in the first three lists, on the contrary, people with Alzheimer's disease or Mild Cognitive Impairment showed lower scores in the last two lists (Yubero et al., 2011). Few neuropsychological instruments explore the involvement of EF in memory failures, improving the number of assessing tools is crucial in clinical domains. Furthermore, the TMS proposes to evaluate executive or amnestic deficits through an immediate recall, guaranteeing a quicker neuropsychological evaluation given the absence of a delayed recall. The design of a neuropsychological rehabilitation program depends on a good diagnosis procedure. If specific deficits (executive or memory) are adequately identified with a neuropsychological test, this favors the selection of the cognitive abilities to be rehabilitated. However, the potential application of TMS also concerns research practice, a possible use of this tool in the future could be implemented in double-dissociation paradigms, to clarify the role of EF in immediate recall. The first use of TMS to distinguish between healthy subjects and patients with memory impairment from the Italian population was presented by Vaccaro et al. (2022). Their findings suggested that the tool could support diagnosis, but the results from the Receiver Operation Curves indicated low discriminate power of EF sub-scale, most likely due to the excessive difficulty of the items (words) that led to no difference between groups. This result is pertinent to the parameters estimated through our multi-group CFA, indeed Italians displayed lower intercepts than other countries. Given this, future studies to address the lack of scalar invariance across countries and consequently modify the structure of TMS are needed to ensure a correct clinical interpretation of TMS outcomes.

Fernandes et al. (2018) found differences between aging and education attainment groups in the TMS's outcomes, as expected, between TMS and age was found significant negative relationship and vice versa with education. These results imply that to guarantee an accurate interpretation of TMS scores independent of socio-demographic features, normative data must be constructed. Indeed, the future studies' aims could concentrate on creating a cut-off for each sub-scale correcting by normative data and testing the other psychometric proprieties of TMS as the criterion validity, for example, through contrasting the scores of TMS to encephalic morphometric measures.

The major problem found in the current study is the lack of scalar invariance across countries, regarding the building of normative data, this issue must be addressed to guarantee an absence of systematic error in the TMS. Nevertheless, the changing of words that compose the lists is probably the solution to the non-invariance of intercepts across countries and is needed for the Italian population before creating the normative data. In summary, TMS could partially avoid the *task impurity problem* by reducing the EF components in the recall task, despite our overall finding suggesting that TMS is adequate for cross-cultural validation, future studies on word frequency of occurrence in the Italian population are needed. Furthermore, testing measurement invariance between patients and healthy subjects is needed to search if a lack of *scalar invariance* remains between these two groups.

5. Limitations and future directions

The first limitation of this study is the restricted number of neuropsychological tests in common between the three samples. We do not have enough information to evaluate the convergent validity of M and EFM. Also, for divergent validity, we considered only the sum score of the Mini-Mental State Examination (MMSE), a composite score of many cognitive functions used in neuropsychological assessment as an overall screening tool.

The second limit of this study is the sample size, the number of participants recommended in Structural equation modeling for *measurement invariance* is >200 for each group.

Another limitation is the Portuguese sample's distribution of age, despite we found some counterintuitive points (the Portuguese participants displayed higher scores than the other groups) was a convenience sample of elderly people, so it reflects a restricted proportion of the population.

Finally, the TMS displayed a lack of scalar invariance across countries,

6. Conclusion

The present study evaluated structural validity and measurement invariance according to three countries of TMS. The results suggested that the structure of TMS is based on three factors (EF, EFM, and M). Furthermore, TMS *configural* and *metric invariance* according to country groups were tenable.

TMS was not *scalar* invariant to country groups. Given this, we recommend meticulous analysis of word frequency occurrence before using TMS in clinical validation paradigms.

Funding

No funds, grants, or other support was received.

Ethics statement

These studies have been approved by "Conselho de Administração do Centro Hospitalar de S. João – EPE". All procedures were carried out following the Declaration of Helsinki, and all participants gave their informed consent.

CRediT authorship contribution statement

Giorgini Roberto: Writing – original draft, Software, Methodology, Formal analysis, Data curation, Conceptualization. **Maestu Fernando:** Writing – review & editing, Visualization, Supervision, Project administration, Methodology, Investigation, Data curation, Conceptualization. **Fernandes Sara Margarida:** Writing – review & editing, Supervision, Investigation, Data curation. **Pastore Massimiliano:** Supervision, Formal analysis, Data curation. **Abellan Maria:** Investigation, Data curation. **Quattrone Andrea:** Visualization, Methodology. **Caparello Sara:** Visualization, Methodology. **Quattrone Aldo:** Writing – review & editing, Supervision. **Vaccaro Maria Grazia:** Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Project administration, Methodology, Investigation, Data curation, Conceptualization.

Declaration of competing interest

None of the co-authors have financial or other conflicts of interest.

Data availability

Raw data and analysis script are available from the corresponding author and/or the last name (MGV) upon request.

References

Baddeley, A. (1992). Working Memory. *Science, 255*(5044), 556–559. https://doi.org/10.1126/science.1736359

Baggetta, P., & Alexander, P. A. (2016). Conceptualization and operationalization of executive function. *Mind, Brain, and Education, 10*(1), 10–33. https://doi.org/10.1111/mbe.12100

Bianchi, A., & Dai Prà, M. (2008). Twenty yeras after Spinnler and Tognoni: New instruments in the Italian neuropsychologist's toolbox. *Neurological Sciences, 29*, 209–217. https://doi.org/10.1007/s10072-008-0970-x

Busch, R. M., McBride, A., Booth, J. E., Vanderploeg, R. D., Curtiss, G., & Duchnick, J. J. (2005). Role of executive functioning in verbal and visual memory. *Neuropsychology, 19*(2), 171–180. https://doi.org/10.1037/0894-4105.19.2.171

Cavaco, S., Gonçalves, A., Pinto, C., Almeida, E., Gomes, F., Moreira, I., … Teixeira-Pinto, A. (2013). Semantic fluency and phonemic fluency: Regression-based norms for the portuguese population. *Archives of Clinical Neuropsychology, 28*(3), 262–271. https://doi.org/10.1093/arclin/act001

Chang, Y. L., Jacobson, M. W., Fennema-Notestine, C., Hagler, D. J., Jennings, R. G., Dale, A. M., & McEvoy, L. K. (2010). Level of executive function influences verbal memory in amnestic mild cognitive impairment and predicts prefrontal and posterior cingulate thickness. *Cerebral Cortex, 20*(6), 1305–1313. https://doi.org/10.1093/cercor/bhp192

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

Craik, F. I. M., Eftekhari, E., Bialystok, E., & Anderson, N. D. (2018). Individual differences in executive functions and retrieval efficacy in older adults. *Psychology and Aging, 33*(8), 1105–1114. https://doi.org/10.1037/pag0000315

Duff, K., Schoenberg, M. R., Scott, J. G., & Adams, R. L. (2005). The relationship between executive functioning and verbal and visual learning and memory. *Archives of Clinical Neuropsychology, 20*(1), 111–122. https://doi.org/10.1016/j.acn.2004.03.003

Fernandes, S. M. (2013). *Teste de Cores e Palavras de Stroop [Stroop color and word test]*. Lisboa: Cegoc- TEA.

Fernandes, S. M., Araújo, A. M., Vázquez-Justo, E., Pereira, C., Silva, A., Paul, N., … Maestú, F. (2018). Effects of aging on memory strategies: A validation of the Portuguese version of the Test of Memory Strategies. *The Clinical Neuropsychologist, 32*, 133–151. https://doi.org/10.1080/13854046.2018.1490456

Folstein, M. F., Folstein, S. E., & McHugh, P. R. (1975). "Mini-mental state". A practical method for grading the cognitive state of patients for the clinician. *Journal of Psychiatric Research, 12*(3), 189–198. https://doi.org/10.1016/0022-3956(75)90026-6

Fossati, P., Amar, G., Raoux, N., Ergis, A. M., Allilaire, J. F., & Francé, F. F. (1999). Executive functioning and verbal memory in young patients with unipolar depression and schizophrenia. *Psychiatry Research, 89*, 171–187. https://doi.org/10.1016/S0165-1781(99)00110-9

García-Laredo, E., Castellanos, M.Á., Badaya, E., Paúl, N., Yubero, R., Maestú, F., … Chacón, J. (2021). Executive functions influence on memory process in patients with paranoid schizophrenia and bipolar disorders with and without psychotic symptoms. A pilot study. *The Spanish Journal of Psychology, 24*, E40. https://doi.org/10.1017/SJP.2021.38

Glisky, E. L., Polster, M. R., & Routhieaux, B. C. (1995). Double dissociation between item and source memory. *Neuropsychology, 9*(2), 229–235. https://doi.org/10.1037/0894-4105.9.2.229

Gregorich, S. E. (2006). Do self-report instruments allow meaningful comparisons across diverse population groups? Testing measurement invariance using the confirmatory factor analysis framework. *Medical Care, 44*, 78–94. https://doi.org/10.1097/01.mlr.0000245454.12228.8f

Higginson, C. I., King, D. S., Levine, D., Wheelock, V. L., Khamphay, N. O., & Sigvardt, K. A. (2003). The relationship between executive function and verbal memory in Parkinson's disease. *Brain and Cognition, 52*(3), 343–352. https://doi.org/10.1016/S0278-2626(03)00180-5

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal, 6*, 1–55. https://doi.org/10.1080/10705519909540118

Levy, G., Jacobs, D. M., Tang, M. X., Côté, L. J., Louis, E. D., Alfaro, B., … Marder, K. (2002). Memory and executive function impairment predict dementia in Parkinson's disease. *Movement Disorders, 17*, 1221–1226. https://doi.org/10.1002/mds.10280

Lobo, A., Ezquerra, J., Burgada, F. G., Sala, J. M., & Seva, A. (1979). El Mini-Examen Cognoscitivo. *Actas Luso-Españolas de Neurología y Psiquiatría, 7*, 189–202.

Magni, E., Binetti, G., Bianchetti, A., Rozzini, R., & Trabucchi, M. (1996). Mini-mental state examination: A normative study in Italian elderly population. *European Journal of Neurology, 3*(3), 198–202. https://doi.org/10.1111/j.1468-1331.1996.tb00423.x

Miyake, A., & Friedman, N. P. (2012). The nature and organization of individual differences in executive functions: Four general conclusions. *Current Directions in Psychological Science, 21*(1), 8–14. https://doi.org/10.1177/0963721411429458

Miyake, A., Friedman, N. P., Emerson, M. J., Witzki, A. H., Howerter, A., & Wager, T. D. (2000). The unity and diversity of executive functions and their contributions to complex "frontal lobe" tasks: A latent variable analysis. *Cognitive Psychology, 41*(1), 49–100. https://doi.org/10.1006/cogp.1999.0734

Morgado, J., Rocha, C., Maruta, C., Guerreiro, M., & Martins, I. (2009). New normative values of mini-mental state examination. *Sinapse, 9*(2), 10–16.

Orsini, A., & Pezzuti, L. (2013). *WAIS-IV. Contributo alla taratura italiana (16–69) (WAIS-IV, contribution to the Italian standardization, ages 16–69)*. Firenze: Giunti OS.

Orsini, A., & Pezzuti, L. (2015). *WAIS-IV. Contributo alla taratura italiana (70–90 anni). (WAIS-IV, contribution to the Italian standardization, ages 16–69)*. Firenze: Giunti OS.

Postle, B. R., Berger, J. S., Esposito, M., & D.. (1999). Functional neuroanatomical double dissociation of mnemonic and executive control processes contributing to working memory performance. *PNAS, 96*, 12959–12964. https://doi.org/10.1073/pnas.96.22.12959

R core Team. (2021). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org.

Rocha, A., Ferreira, C., Barrete, H., Moreira, A., & Machado, M. (2008). *WAIS–III- Escala de Inteligência de Wechsler para Adultos[WAIS-III-Wechsler Adults Intelligence Scale]*. Lisboa: CEGOC-TEA.

Rönkkö, M., & Cho, E. (2022). An updated guideline for assessing discriminant validity. *Organizational Research Methods, 25*(1), 6–14. https://doi.org/10.1177/1094428120968614

Roodenrys, S., Lethbridge, A., Hinton, M., Nimmo, L. M., & Hulme, C. (2002). Word-frequency and phonological-neighborhood effects on verbal short-term memory. *Journal of Experimental Psychology: Learning Memory and Cognition, 28*(6), 1019–1034. https://doi.org/10.1037/0278-7393.28.6.101

Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software, 48*(2), 1–36. https://doi.org/10.18637/jss.v048.i02

Scarpina, F., & Tagini, S. (2017). The stroop color and word test. *Frontiers in Psychology, 8*. https://doi.org/10.3389/fpsyg.2017.00557 (1664-1078).

Stroop, J. R. (1935). Studies of interference in serial verbal reactions. *Journal of Experimental Psychology, 18*(6), 643–662. https://doi.org/10.1037/h0054651

Vaccaro, M. G., Liuzza, M. T., Pastore, M., Paúl, N., Yubero, R., Quattrone, A., … Maestú, F. (2022). The validity and reliability of the Test of Memory Strategies among Italian healthy adults. *PeerJ, 10*. https://doi.org/10.7717/peerj.14059

Wagenmakers, E. J., & Farrell, S. (2004). AIC model selection using Akaike weights. *Psychonomic Bulletin & Review, 11*, 192–196. https://doi.org/10.3758/BF03206482

Wechsler, D. (1981). *Wechsler adult intelligence scale-revised (WAIS-R)*. Psychological Corporation.

Wechsler, D., de la Guía, E., & Vallar, F. (2012). *WAIS-IV: escala de inteligencia de Wechsler para adultos-IV*. Madrid: Pearson.

Whiteside, D. M., Kealey, T., Semla, M., Luu, H., Rice, L., Basso, M. R., & Roper, B. (2016). Verbal fluency: Language or executive function measure? *Applied Neuropsychology: Adult, 23*(1), 29–34. https://doi.org/10.1080/23279095.2015.1004574

Yubero, R., Gil, P., Paul, N., & Maestu, F. (2011). Influence of memory strategies on memory test performance: A study in healthy and pathological aging. *Aging, Neuropsychology, and Cognition, 18*(5), 497–515. https://doi.org/10.1080/13825585.2011.597840