EVALUAR

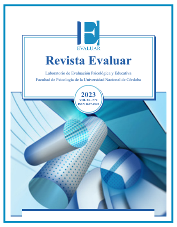
Revista Evaluar

Laboratorio de Evaluación Psicológica y Educativa
Facultad de Psicología de la Universidad Nacional de Córdoba

2023
VOL 23 - N°2
ISSN 1667-4545

2023, Vol. 23, No. 2
ISSN 1667-4545
Recuperado de https://revistas.unc.edu.ar/index.php/revaluar

Laboratorio de Evaluación Psicológica y Educativa
Facultad de Psicología - Universidad Nacional de Córdoba



Suicidal Behavior at Work Scales: Development and Validation of the Work-Related Suicidal Ideation, Defeat, and Entrapment Brief Scales

Escalas de comportamiento suicida en el trabajo: Desarrollo y validación de las escalas breves de ideación suicida, derrotismo y atrapamiento relacionado con el trabajo

Lillian V. Rovira-Millán [1], Rafael A. Blanco-Rovira [2], Ana C. López-Iglesias [1], and Ernesto Rosario-Hernández [3, 4] *

1 - University of Puerto Rico, Cayey, Puerto Rico.
2 - Carlos Albizu University, San Juan, Puerto Rico.
3 - School of Behavioral & Brain Sciences, Ponce Health Sciences University, Ponce, Puerto Rico.
4 - Ponce Research Institute, Ponce Health Sciences University, Ponce, Puerto Rico.

Recibido: 16/02/2023 **Revisado:** 01/03/2023 **Aceptado:** 07/03/2023

Abstract

Background: Suicide is a health problem around the world, since suicide rates among Americans aged 45 to 54 is the highest, and most of these individuals were employed at the time of their death. Thus, there is a need to better understand suicidal behavior at work by developing appropriate measurement instruments in order to create prevention and treatment programs. Therefore, the aim of this study was to develop and validate three brief self-report measures of suicidal behavior at work: defeat, entrapment, and work-related suicidal ideation.
Materials and Methods: A total of 1,829 employed individuals from different organizations in Puerto Rico participated in this cross-sectional research design. We conducted item, exploratory, and confirmatory factor analyses. Also, we tested measurement invariance of the new brief scales of suicidal behavior at work.
Results: The final version of the suicidal behavior at work brief scales obtained excellent reliability coefficients using Cronbach's alpha and McDonald's omega techniques. The results of the EFA and CFA support their internal structure. The new scale appears to be invariant among groups.
Conclusion: The scores of the new suicidal behavior at work brief scales appear to be reliable, valid, and invariant, which will help to study and to better understand these behaviors in order to create treatments and prevention programs in our workplaces.

Keywords: *CFA, EFA, Defeat, Entrapment, Measurement Invariance, Suicidal Ideation, Work-Related Suicidal Ideation*

Resumen

Antecedentes: El suicidio es un problema de salud en todo el mundo, ya que las tasas de suicidio entre los estadounidenses de 45 a 54 años son las más altas, y la mayoría de estas personas estaban empleadas en el momento de su muerte. Por lo tanto, existe la necesidad de comprender mejor el comportamiento suicida en el trabajo desarrollando instrumentos de medición apropiados para así crear programas de prevención y tratamiento. Por lo tanto, el objetivo del presente estudio fue desarrollar y validar tres medidas breves de autoinforme de comportamiento suicida en el trabajo: la percepción de derrotismo, el atrapamiento y la ideación suicida relacionada con el trabajo.
Materiales y Método: Un total de 1,829 personas empleadas de diferentes organizaciones en Puerto Rico participaron en este diseño de investigación transversal. Realizamos análisis de reactivos, factores exploratorios y confirmatorios. Además, probamos la invariancia de medición de las nuevas escalas breves de comportamiento suicida en el trabajo por género, edad, entre otros.
Resultados: La versión final de las escalas breves de conducta suicida en el trabajo obtuvo excelentes coeficientes de confiabilidad mediante las técnicas alfa de Cronbach y omega de McDonald. Los resultados de los análisis de factores exploratorios y confirmatorios respaldan su estructura interna. Las nuevas escalas parecen ser invariantes.
Conclusiones: Las puntuaciones de las nuevas escalas breves de conducta suicida en el trabajo parecen ser confiables, válidas e invariantes, lo que ayudará a estudiar y comprender mejor estas conductas para crear tratamientos y programas de prevención en los lugares de trabajo.

Palabras clave: *CFA, EFA, fracaso, atrapamiento, invariancia de medida, ideación suicida, ideación suicida relacionada con el trabajo*

***Correspondence to:** Dr. Ernesto Rosario-Hernández, School of Behavioral and Brain Sciences, Clinical Psychology Programs, Ponce Health Sciences University, PO Box 7004, Ponce, Puerto Rico, 00732-7004. Phone: 787-840-2575, Ext. 5512. E-mail: erosario@psm.edu
Author's note: LVR-M: lillian.rovira@upr.edu, https://orcid.org/0000-0001-7479-7401; RAB-R: rb0230039@sju.albizu.edu, https://orcid.org/0000-0002-8153-1967; ACL-I: ana.lopez23@upr.edu, https//orcid.org/0000-0001-7101-6745; ER-H: erosario@psm.edu, https//orcid.org/0000-0002-2523-6162.

How to cite: Rovira-Millán, L. V., Blanco-Rovira, R. A., López-Iglesias, A. C., & Rosario-Hernández, E. (2023). Suicidal Behavior at Work Scales: Development and Validation of the Work-Related Suicidal Ideation, Defeat, and Entrapment Brief Scales. *Revista Evaluar*, 23(2), 1-18. Retrieved from https://revistas.unc.edu.ar/index.php/revaluar
Participaron en la edición de este artículo: Marianela López, Stefano Macri, Eugenia Barrionuevo, Pablo Carpintero, Agustina Mangieri, Florencia Ruiz, Benjamín Casanova, Jorge Bruera.

Introduction

No matter how industrialized or wealthy a nation is, suicide is one of the most significant health and behavioral problems (Otsuka et al., 2016). The World Health Organization (WHO, 2021) estimates that 800,000 individuals worldwide commit suicide each year, making it the third global leading cause of death. According to Mortali and Moutier (2019), who rate suicide as the tenth-leading cause of death overall and the fourth-leading cause for people under the age of 65, it is also a significant public health concern in the United States (US). In Puerto Rico, according to the Commission for the Prevention of Suicide (CPS, 2016), a suicide happens every 28 hours, or at least once every day.

The Morbidity and Mortality Weekly Report (MMWR) asserts that this critical health concern affects workplaces as well (Peterson et al., 2018). The suicide rate among Americans in their working years climbed by 34% between 2000 and 2016. However, unlike a workplace injury, a suicide that takes place at work does not count as “occupational suicide” (Kasl & Jones, 2003). Tiesman et al. (2015) found that suicide rates have sharply increased recently, even though national workplace suicide trends have not been widely studied. According to Mortali and Moutier (2019), the suicide rate among Americans aged 45 to 54 is the highest (19.72 per 100,000) and most of these individuals were employed at the time of their deaths.

The research of work-related suicide behavior involves studying its relationship with some aspects that have been considered as possible predictors of suicide behavior such as previous suicidal attempts, depression, hopelessness and mental disorders (e.g., O’Connor & Nock, 2014). Even though suicide behaviors are public health issues, there seems to be a paucity of empirical research testing the strength, direction, and nature of these relationships at work. Therefore, research efforts are needed to understand the etiology of suicide (e.g., Suominen et al., 2004), and especially, to assess suicidal ideation and better manage suicide behavior (Avendaño-Prieto et al., 2018). Moreover, the rising number of workplace suicides highlights the need for more study of occupation-specific risk factors and the development of evidence-based initiatives that may be applied in the workplace (Tiesman et al., 2015). However, to increase these research efforts, it is important to develop measurement instruments of these suicide behavior predictors.

There are new theoretical frameworks which focus on suicidal behavior from ideation to action (Klonsky et al., 2018). One of these theories under this paradigm is the Integrated Motivational-Volitional (IMV) model of suicidal behavior proposed by O’Connor (2011). O’Connor combined the primary components of the most popular models of suicidal conduct into the IMV model of suicidal behavior, an integrated three-phase model that aims to distinguish between suicide ideators and suicide attempters. The IMV model is a three-phase framework to elucidate the origins of suicidal ideation and behavior, which are pre-motivational, motivational, and volational. Background elements and triggering events are included in the pre-motivational phase. This pre-motivational phase is significant because it emphasizes how the interacting diathesis-environment-life-events triad that makes up this phase of the model influences the IMV model. In other words, suicide ideation and conduct are the outcome of a biological or genetic interaction that confers a susceptibility that is activated or increased in the presence of stress. Suicidal ideation and intention development, on the other hand, are part of the motivational phase. Feelings of defeat can lead to feelings of entrapment, which can result in suicide ideation and intent. Self-moderators like rumination can contribute to the move from concepts of defeat to feelings of entrapment. The volitional phase describes when suicide attempts are more likely to occur. According to the IMV model, a set of elements, known as volitional moderators, influences the conditions and situations in which a person is more likely to engage in suicidal behavior. A volitional moderator, according to O'Connor, is any factor that bridges the suicidal ideation-behavior gap, that is to say, any element that makes it probable that people will act on their suicidal ideation (e.g., impulsivity).

Research Purpose

The purpose of this study was to develop and validate three brief self-report measures of suicidal behavior related to work (work-related suicidal ideation, feelings of defeat and entrapment) based on the IMV model of suicide behavior. Moreover, another objective was to examine whether these new scales were invariant in terms of gender, age, job position, type of organization and type of contract.

Method

Participants

A total of 1,829 protocols of employed individuals from different organizations in Puerto Rico that had participated in two studies conducted by the authors were used in this instrumental research design. In those two studies, they were selected based on their availability and volition. Besides, anonymity and the right to abandon the research were guaranteed when they considered it necessary. Table 1 shows the description of the sample's sociodemographic characteristics.

Materials

Background questionnaire. We designed a background questionnaire to gather information about the participants in the research. In this background questionnaire, we asked the participants to provide information about their gender, age, tenure, marital status, among others, to enable us to describe the subjects of the study.

Suicidal ideation. To measure suicidal ideation, we developed the Work-Related Suicidal Ideation Scale (WRSIS). The WRSIS is composed of 15 items, which intent to measure suicidal ideations related to work issues. This instrument is in a Likert-frequency response format ranging from 1 (*Never*) to 6 (*Always*).

Defeat. We developed the Defeat Scale to measure feelings of defeat. This is a six-item instrument in a Likert-agreement response format ranging from 1 (*Totally Disagree*) to 6 (*Totally Agree*), which aims to measure general feelings of defeat.

Entrapment. We developed the Entrapment Scale to measure feelings of being trapped without possibilities to get out of a situation. This is a six-item instrument in a Likert-agreement response format ranging from 1 (*Totally Disagree*) to 6 (*Totally Agree*), which intends to measure general feelings of entrapment.

Depression. To measure depression, we used the PHQ-9 developed by Kroenke et al. (2001). The PHQ-9 is a nine-item questionnaire used for the assessment of depressive symptoms in primary care settings. This questionnaire assesses the presence of depressive symptoms over the 2 weeks prior to the test's being filled out. Each of the items can be scored from 0 (*not at all*) to 3 (*nearly every day*).

Anxiety. To measure anxiety, we used the GAD-7 (Spitzer et al., 2006). The GAD-7 is a seven-item questionnaire that quantifies general anxiety symptomatology and by which patients were asked how often, during the prior 2 weeks, they

Table 1
Socio Demographic Characteristics of the Sample.

Variable	Freq.	%	Variable	Freq.	%
Gender			**Type of Employment**		
Males	731	40.0	Tenure	1,217	66.5
Females	1,026	56.1	Temporary	549	30.0
Age			**Years Working**		
21-30 (Early Career)	378	20.7	1 - 5	669	36.6
31-50 (Peak of Career)	313	17.1	6-10	321	17.6
51 (Past Peak of Career)	275	15.0	11-15	257	14.1
Marital Status			16-20	181	9.9
Single	713	39.0	21-25	155	8.5
Married	688	37.6	26-30	106	5.8
Widowed	50	2.8	31	97	5.3
Divorced	166	9.1	**Type of Organization**		
Living Together	198	10.9	Public	576	31.9
Job Position			Private	1,080	59.0
Managerial	351	19.2		**Mean**	**SD**
Non-Managerial	1,409	77.0	**Education** (In Years)	15.20	2.80

Note. n = 1,829.

were bothered by each symptom. Response options were *not at all*, *several days*, *more than half the days*, and *nearly every day*, scored as 0, 1, 2, and 3, respectively.

Rumination. We used the Affective Rumination subscale of Work-Related Rumination Scale-Spanish version (Cropley et al., 2012; Rosario-Hernández et al., 2021) to measure one of the moderators of the IMV model of suicide behavior (O'Connor et al., 2011). As part of the current study, we used only the Affective rumination subscale of the WRRS-Spanish version.

Social desirability. We used the Social Desirability Scale developed by Rosario-Hernández and Rovira-Millán (2002). This is an 11-item instrument in a Likert-agreement response format ranging from 1 (*Totally Disagree*) to 6 (*Totally Agree*), which intends to measure a response bias in which people respond to a test thinking what is socially acceptable.

Procedures

The Institutional Review Board (IRB) of the Ponce Health Sciences University socially approved of the realization of the two studies. Protocol numbers were 160913-ER and 180313-ER.

In order to develop the instruments, we revised the literature and other similar measures. Thus, we developed 15 items for the WRSI, and six items for each of the Defeat Scale, and the Entrapment Scale. The developed items of the scales were administered to a sample of employees from different organizations in Puerto Rico. We conducted individualized item analysis for each of the three scales. It was established as criterion following recommendation of some of the literature (e.g., DeVellis, 2017; Spector, 1992) that all items with an item-total correlation or r_{bis} $\geq .30$ were included in the next step of the exploratory factor analysis (EFA). Also, EFA at first was conducted individually for each scale and the criterion established in the EFA was that all items with a factor loading ≥ .30 on its corresponding factor were selected (e.g., Kline, 1994). After conducting all individualized scale's EFA, we conducted an EFA including all the items of the three scales and then we proceeded to conduct a confirmatory factor analysis (CFA) using structural equation modeling via the lavaan package of the R program (Rosseel, 2012). Moreover, to establish convergent and divergent validity, we correlated observed scores and latent constructs of the three new scales and the Social Desirability Scale (Rosario-Hernández & Rovira-Millán, 2002). Finally, reliability and descriptive statistics were computed for the new scales.

Data analysis

First, we performed descriptive statistics analyses to obtain sociodemographic characteristics of the sample. Also, we conducted descriptive analyses of the three scale's items, such as the mean and standard deviation values. An item analysis was also performed to obtain the discrimination index which is also known as corrected item-total correlation or r_{bis}. We used the whole sample to perform these descriptive and item analyses.

Second, the total sample was randomly split into two samples, and then each of them was also randomly split into two more samples each hereafter referred to as sample 1 (n_1), sample 2 (n_2), sample 3 (n_3), and sample 4 (n_4). This method allows examining the stability of the structural factor's solution across the halves (Fabrigar et al., 1999). Third, exploratory factor analyses (EFAs) were conducted with sample 1 and sample 2 using SPSS v.28 (IBM, 2021). EFAs were conducted using the extraction method of principal axis factoring with a direct oblimin rotation. As selection criteria, all those items that obtained a factor loading ≥ .30 in the factor to which it supposedly belongs were selected as recommended by Kline (1994). At first, we individualized EFAs to each set of items of each scale and then we included all the items of the three scales that comply with the criteria and conducted another EFA with all of them using sample 1. In order to cross-validate the three-factor structure, we conducted another EFA using sample 2.

Fourth, all items selected from the EFA were subjected to CFA using the structural equation modeling to examine the internal structure of the suicidal behaviors at work brief scales using the weighted least squares-mean and variance adjusted (WLSMV) estimator with the lavaan package of the R3.6.3 program (Rosseel, 2012), which robustly deals with potentially non-normal data and items are treated as ordinal (Li, 2016a, 2016b). To evaluate the results of the CFA, several fit indices of the structural equation models were used. Kline (2016) recommends the use of at least four fit indices, although more can be reported. One of the indices reported is Chi-Square (χ^2), which is a fundamental index of absolute fit and is basically the same one that is used when you want to examine the association between nominal variables. However, the crucial difference when it is used as an index of fit in the structural equations model is that the researcher looks for no differences between the matrices to support that the tested model represents the data (Hair et al., 2019). Given the fact that the χ^2 is sensitive to the sample size and, therefore, the probability of rejecting the hypothesized model increases when the sample size grows, it is recommended to take into account other indices (Marsh et al., 1996). This way, the Root Mean Square Error of Approximation (RMSEA; Byrne, 2016; Hu & Bentler, 1999) was used, in which values ≤ .05 indicate a good fit of the model, values < .08 for the RMSEA indicate an acceptable fit; values ranging from .08 to .10 are considered as mediocre (Browne & Cudeck, 1993; MacCallum et al., 1996). In addition, Standardized Square Root Mean Residual (SRMR; Hu & Bentler, 1995) was used, which examines the average difference between predicted and observed variances and covariances, based on the residual standard error. The lower the SRMR, the better the fit of the model and, to be considered an acceptable model, it must be equal to or less than .05. On the other hand, the Bentler Comparative Fit Index (CFI) was used as an increased fit index to compare the theoretical model with the null model, which assumes that the latent variables of the model they do not correlate with each other and values greater than .90 are considered acceptable (Hair et al., 2019). Another increased adjustment index is the Tucker-Lewis Index (TLI), which reflects the proportion in which the theoretical model improves the adjustment in relation to the null model (Littlewood-Zimmerman & Bernal-García, 2011; Tucker & Lewis, 1973). Values greater than .90 are considered acceptable. We conducted CFA's with sample 3 to calibrate and sample 4 to validate results.

Fifth, we recombined the samples and assessed measuring invariance across gender, age, job position, type of organization, and type of contract. We tested configural invariance, metric invariance, and scalar invariance as suggested by some in the literature (e.g., Byrne, 2016; Muthén & Muthén, 2012; Wang & Wang, 2012). We conducted hierarchical tests for invariance of measurement parameters. First, we examined the configural invariance model or pattern invariance, which imposes no equality restrictions on model parameters. This is a necessary condition for testing invariance by comparing it with other invariance models based on fit indices. Second, we examined the weak invariance model or metric invariance. In this model, the factor loadings are treated as invariant across groups. This ensures that the measures are on the same scale across groups. Third, we examined the strong invariance model. This model imposes invariance on both factor loadings and item intercept across groups. This is to ensure that the underlying factors can be compared across groups. We capitalized on fit index differences for CFI, SRMR, and RMSEA (i.e., ΔCFI, ≤-.01, ΔSRMR & ΔRMSEA ≥.015) reference points as recommended by Chen (2007), who found in a Monte Carlo study that these indices were equally sensitive to all types of invariances. Notably, as the χ^2 is known to be highly influenced by the sample size (e.g., Rigdon, 1995), it was reported but not considered as fit index for the invariance testing.

Sixth, with the recombined sample, we examined the convergent and divergent validity of the three new self-report measures of suicidal behavior at work by their covariation and estimating the average variance extracted (AVE), maximum shared variance (MSV), and the shared mean variance (ASV) based on a CFA with the total sample. According to Fornell-Larcker (1981), as the value of the AVE is greater than .50, it implies that it measures more variance of the construct and less error. Furthermore, if all AVE constructs are higher than .50 and are higher than the MSV and ASV, it supports the convergent and divergent validity of the scales. Similarly, we assessed

Table 2
Descriptive statistics and corrected item-total correlation (r_{bis}) of the Three Self-Report Work-Related Suicidal Behavior Brief Scales.

Item	Mean	SD	r_{bis}	Item	Mean	SD	r_{bis}
Work-Related Suicidal Ideation				**Defeat**			
WRSI-1	1.20	0.644	.675	Def-1	2.16	1.573	.130
WRSI-2	1.14	0.545	.731	Def-2	1.47	1.111	.664
WRSI-3	1.16	0.591	.712	Def-3	1.72	1.369	.594
WRSI-4	1.17	0.598	.632	Def-4	1.41	1.045	.661
WRSI-5	1.10	0.492	.802	Def-5	1.34	0.961	.670
WRSI-6	1.07	0.387	.848	Def-6	1.25	0.855	.684
WRSI-7	1.06	0.408	.852	**Entrapment**			
WRSI-8	1.06	0.375	.822	Ent-1	1.32	0.952	.799
WRSI-9	1.05	0.375	.820	Ent-2	1.31	0.917	.776
WRSI-10	1.06	0.381	.820	Ent-3	1.29	0.868	.817
WRSI-11	1.06	0.391	.847	Ent-4	1.38	1.013	.808
WRSI-12	1.11	0.482	.798	Ent-5	1.66	1.308	.753
WRSI-13	1.07	0.394	.831	Ent-6	1.82	1.463	.623
WRSI-14	1.06	0.394	.817				
WRSI-15	1.06	0.381	.824				

Note. n = 1,829; SD = Standard Deviation; r_{bis}= item-total correlation.

the convergent and divergent validity of the new scales by correlating observed scores of the scales with each other and with observed scores from other instruments measuring rumination, depression, anxiety, and social desirability. Finally, we performed internal consistency reliability via Cronbach's alpha and McDonald's omega, standard error of measurement and 95% confidence interval and descriptive statistics to estimate mean and standard deviation of the scales.

Results

First, we obtained descriptive statistics and conducted an analysis of the items from the three suicidal behavior brief scales. Table 2 shows the mean, the standard deviation, and the corrected item-total correlations (r_{bis}). Only item 1 of the Defeat Scale did not reach a r_{bis} of .30; therefore, it was eliminated and not included in subsequent analyses.

EFA were performed for each scale individually with sample 1. The results of these EFA for the defeat and entrapment scales showed a one-dimensional internal structure, while the WRSI Scale showed a two-factor structure. When examining the items, those that expressed work-related suicidal ideation loaded on Factor 1 and those that expressed suicidal ideation, in general, loaded on Factor 2. Therefore, it was decided to select those items that expressed work-related suicidal ideation, that is, the items that loaded on Factor 1. Thus, the nine items from the WRSI

Scale, five from the Defeat Scale and six from the Entrapment Scale, were included in the next EFA and the results showed an internal structure of three-factors explaining 69.55% of the variance. All items obtained factor loadings ≥ .30 on their respective factors as suggested by the literature (e.g., Kline, 1994); however, item 5 of the Defeat Scale had cross-loading on Factor 2 and 3. Nevertheless, it was included in subsequent analyses because it obtained a much higher loading in its respective Factor 3 and barely passed the threshold of .30 on Factor 2 (see Table 3). A second EFA was conducted, but this time with sample 2 to cross-validate results from previous EFA. As shown in Table 3, the three-factor structure was also supported and explained 74.72% of the variance.

We tested three competitive models, one-factor, two-factor, and three-factor structure models of the suicidal ideation behaviors at work scales using structural equation modeling. We used sample 3 for this first CFA as a calibration sample. The one-factor model included all items of the three scales loading just in one factor and obtained acceptable fit indices, except for the SRMR since it exceeded the recommended threshold of .05 (Hu & Bentler, 1995). Also, we tested a two-factor model in which items from the Defeat and Entrapment Scales as one factor as some findings in the literature suggest (e.g., Taylor, Wood et al., 2010) and WRSI items as another one. This two-factor model obtained better fit indices than the one-factor model, including a better SRMR although still above the threshold of .05 (see Table 4). Finally, we tested the three-factor model and results showed that this was the best fitted model of all, since all fit indices were within the thresholds (see Table 4). Thus, as fit indices of the three-factor model were very good, it was decided to probe this model with sample 4 to cross validate the three-factor structure model. All fit indices obtained were very good (see Table 4), supporting the three-factor model implicating that each scale measures a different construct and all items of the three scales obtained factor loadings ≥ .70, except item 6 of the Entrapment Scale with the whole sample, but with sample 3 and sample 4 were above .70 (see Table 5).

Table 4
Fit indices of the Three Self-Reports Work-Related Suicidal Behavior Brief Scales for models tested.

Model	χ^2 (*df*)	SRMR	RMSEA (90% CI)	CFI	TLI
1 Factor (Sample 3)	344.524* (170)	.086	.048 (.040, .055)	.994	.993
2 Factor (Sample 3)	299.107* (169)	.053	.041 (.033, .049)	.996	.995
3 Factor (Sample 3)	245.685* (167)	.047	.032 (.023, .041)	.997	.997
3 Factor (Sample 4)	266.857* (167)	.042	.036 (.028, .044)	.998	.997

Note. n_3 = 454; n_4 = 457; χ^2 = chi-square statistic; *df* = degree of freedom; SRMR = Standardized Root Mean Squared Residual; CFI = Comparative Fit Index; TLI = Tucker-Lewis Index; RMSEA = Root Mean Square Error of Approximation; CI = Confidence Interval.

Measurement invariance was achieved with a bottom-up approach, from an unrestricted model to a model with strong restriction (Stark et al., 2006). Thus, we tested an unrestricted model of equality (configurational invariance) and continued with successive restrictions applied to factor loadings, thresholds (metric invariance) and intercepts (scalar invariance). Considering the sample size (> 300), the invariance criteria were: CFI < .010, SRMR < .030, and RMSEA < .015 (Chen, 2007). As such, measurement invariance in every group analyzed (i.e., gender, age, job position, organization type, and employment type) was

Table 3
Exploratory factor analyses of the Three Self-Report Work-Related Suicidal Behavior Brief Scales on sample 1 and sample 2.

Item	Sample 1				Sample 2			
	Factor			h^2	Factor			h^2
	1	2	3		1	2	3	
WRSI-6	**.78**			.685	**.87**			.862
WRSI-7	**.90**			.838	**.96**			.928
WRSI-8	**.88**			.826	**.98**			.929
WRSI-9	**.93**			.899	**.85**			.747
WRSI-10	**.84**			.790	**.98**			.954
WRSI-11	**.84**			.702	**.86**			.808
WRSI-13	**.80**			.768	**.93**			.849
WRSI-14	**.93**			.826	**.96**			.928
WRSI-15	**.79**			.744	**.87**			.816
Def-2			**.67**	.520			**.68**	.477
Def-3			**.56**	.456			**.63**	.505
Def-4			**.69**	.603			**.79**	.626
Def-5		.31	**.56**	.655			**.79**	.746
Def-6			**.60**	.632			**.90**	.797
Ent-1		**.81**		.848		**.74**		.649
Ent-2		**.74**		.752		**.84**		.744
Ent-3		**.87**		.799		**.84**		.805
Ent-4		**.80**		.678		**.86**		.780
Ent-5		**.72**		.555		**.75**		.584
Ent-6		**.54**		.333		**.64**		.411
Eigen Value	8.88	52.38	52.38		9.63	54.45	54.45	
% Variance Explained	7.87	13.70	66.08		6.96	15.12	69.57	
% Variance Accumulated	5.98	3.47	69.55		7.28	5.15	74.72	
KMO	.909				.915			
χ^2 (*df*)	10,671*	(190)			11,675*	(190)		

Note. $n_1 = 458$; $n_2 = 460$; **p* < .01; *df* = degree of freedom.

Table 5
Factor Loadings of Items of the Three Self-Reports Work-Related Behavior Brief Scales from the Confirmatory Factor Analyses.

Scale	Item	Factor Loadings		
		Sample 3	Sample 4	Total Sample
Work-Related Suicidal Ideation	WRSI-6	.984	.980	.963
	WRSI-7	1.01	.982	.984
	WRSI-8	.974	.952	.973
	WRSI-9	.973	.987	.979
	WRSI-10	.954	.990	.973
	WRSI-11	.966	.986	.973
	WRSI-13	.974	.981	.975
	WRSI-14	.990	1.00	.983
	WRSI-15	.993	.972	.980
Defeat	Def-2	.805	.873	.831
	Def-3	.788	.828	.815
	Def-4	.888	.916	.890
	Def-5	.912	.893	.928
	Def-6	.941	.967	.955
Entrapment	Ent-1	.957	.956	.955
	Ent-2	.943	.966	.956
	Ent-3	.987	.966	.970
	Ent-4	.909	.926	.921
	Ent-5	.885	.881	.876
	Ent-6	.812	.785	.777

Note. n_3 = 454; n_4 = 457; n_T = 1,829.

good and complied with the established criteria. The differences between fit indices (Δ_{CFI}, Δ_{RMSEA}, and Δ_{SRMR}) were within limits, suggesting that the three self-report measures of suicidal behavior were invariant among those groups (see Table 6).

Table 7
Average Variance Extracted (AVE), Maximum Shared Variance (MSV), Average Shared Variance (ASV) and correlation between latent constructs to establish convergent and divergent validity.

Scale	AVE	MSV	ASV	1	2	3
1. WRSI	.95	.64	.63	1		
2. Def	.78	.72	.67	.78**	1	
3. Ent	.83	.72	.68	.80**	.85**	1

Note. n = 1,829; *p <.05; **p <.01.

First, to evaluate convergent validity of reflective construct as work-related suicidal ideation, feelings of defeat and entrapment, we checked that the average variance extracted (AVE) value of items of the three-scales were developed and all were ≥.50 (see Table 7). We calculated the AVE using the whole sample for WRSI, Defeat, and Entrapment Scales and they were .95, .78, and .83, respectively; all well above the threshold of .50 (see Table 7). Also, we estimated the maximum shared variance (MSV) and the average shared variance (ASV) to establish divergent validity and all AVEs of the three new scales were larger than the MSV and the ASV, supporting convergent and divergent validity of the three brief self-report measures of suicidal behavior at work.

In order to establish the convergent and divergent validity of the three new brief scales of suicidal behavior at work, we correlated their observed scores with observed scores of depression, anxiety, rumination, and social desirability measures. Table 8 shows that the observed score correlations between the three brief scales of suicidal behavior at work with depression, anxiety, rumination, and social desirability correlated in the expected direction and magnitude. For example, entrapment scores were higher in terms of depression, anxiety, and rumination, which can be

Table 6
Measurement Invariance of the Three Self-Reports Work-Related Suicidal Behavior Brief Scales by Gender, Age, Job Position, Type Organization, and Type of Employment.

Model	χ^2 (*df*)	SRMR	RMSEA (90% CI)	CFI	Reference Model	Δ^2	ΔSRMR	ΔRMSEA	ΔCFI
Multigroup analysis by gender (male/female)									
1. Configural	497.399* (334)	.044	.023 (.019, .028)	.998	-----	-----	-----	-----	-----
2. Metric	512.284* (351)	.046	.023 (.018, .027)	.998	1	+14.885	+.002	.000	.000
3. Scalar	588.959* (401)	.044	.023 (.019, .027)	.998	2	+76.675	-.002	.000	.000
Multigroup analysis by age (21-30/31-50/51)									
1. Configural	612.959* (501)	.078	.019 (.013, .024)	1.00	-----	-----	-----	-----	-----
2. Metric	641.451* (535)	.079	.018 (.012, .023)	1.00	1	+28.582	+.001	-.001	.000
3. Scalar	753.803* (615)	.078	.019 (.014, .024)	1.00	2	+112.352	-.001	+.001	.000
Multigroup analysis by job position (managerial/non-managerial)									
1. Configural	630.353* (334)	.035	.032 (.028, .036)	.997	-----	-----	-----	-----	-----
2. Metric	604.418* (351)	.037	.029 (.025, .032)	.998	1	-25.935	+.002	-.003	+.001
3. Scalar	701.587* (424)	.034	.027 (.024, .031)	.998	2	+97.169	-.003	-.002	.000
Multigroup analysis by organization type (public/private)									
1. Configural	565.729* (334)	.032	.029 (.025, .033)	.998	-----	-----	-----	-----	-----
2. Metric	597.756* (351)	.035	.029 (.025, .033)	.998	1	+32.027	+.003	.000	.000
3. Scalar	639.324* (423)	.033	.025 (.021, .029)	.998	2	+41.568	-.002	-.004	.000
Multigroup analysis by type of employment (tenure/temporary)									
1. Configural	583.843* (334)	.033	.029 (.025, .033)	.998	-----	-----	-----	-----	-----
2. Metric	616.399* (351)	.037	.029 (.025, .033)	.997	1	+32.556	+.004	.000	-.001
3. Scalar	677.257* (420)	.034	.026 (.023, .030)	.998	2	+60.858	-.003	-.003	+.001

Note. *p < .05; *df* = Degree of Freedom.

Table 9
Descriptive statistics and reliability of the three self-report work-related suicidal behavior brief scales.

Scale	# Items	Mean	SD	Reliability (CI)		sem	95% CI	Min	Max	Possible Range
				∝	ω					
WRSI	9	9.55	3.19	.976 (.963, .983)	.976 (.962, .983)	0.49	±1	9	52	9 – 54
Def	5	7.18	4.36	.865 (.838, .888)	.866 (.840, .888)	1.59	±3	5	30	5 – 30
Ent	6	8.78	5.46	.902 (.886, .918)	.902 (.884, .917)	1.70	±3	6	36	6 – 36

Note. n = 1,829; ∝ = Cronbach's alpha; ω = McDonald's omega; CI = Confidence Interval; sem = Standard Error of Measurement; WRSI = Work-Related Suicidal Ideation; Def = Defeat; Ent = Entrapment.

Table 8
Correlation between observed scores of the Three Brief Self-Reports of Suicidal Behavior at Work and other measures to establish convergent and divergent validity.

Scale	WRSI	Def	Ent
WRSI	1		
Def	.53**	1	
Ent	.59**	.68**	1
Dep	.41**	48**	.57**
Anx	.40**	.47**	.56**
Rum	.22**	.34**	.38**
SD	-.10*	-.04^{NS}	-.13*

Note. n = 898; *p < .05; **p < .01; NS = Not Significant; WRSI = Work-Related Suicidal Ideation; Def = Defeat; Ent = Entrapment; Dep = Depression; Anx = Anxiety; Rum = Rumination; SD = Social Desirability.

considered to have medium to large effect sizes (Cohen, 1988). Finally, observed correlations between scores of the three new developed scales and social desirability scores were very low and close to zero.

Reliability and Descriptive Statistics

We estimated the mean, the standard deviation, the standard error of measurement, and the 95% of confidence interval for the scores of the final version of the three suicidal behaviors at work scales (see Table 8). Moreover, we estimated the reliability using Cronbach's alpha and McDonald's omega with their respective confidence intervals, and all reliability coefficients were above .70 as suggested by some of the literature (e.g., DeVellis, 2017; Spector, 1992).

Discussion

The objective of this study was to develop and validate three brief self-report measures of suicidal behavior at work: (1) WRSI, (2) Feelings of Defeat, and (3) Entrapment Scales. The EFA results that were scale-specifically supported a unidimensional internal structure of each of the three scales. Additionally, when we integrated all three scales' items and performed an EFA, all items loaded onto their respective factors, which allowed to corroborate the internal structure of one-factor for each scale. Further, similar outcomes were obtained when an EFA was conducted using sample 2, which also demonstrated that the set of items for each scale loaded on its corresponding factor. Meanwhile, three models were tested for the CFA: (a) one-factor, in which all items loaded on a single factor; (b) two-factor, in which the items from the WRSI scale loaded on one factor and those from the defeat and entrapment scales loaded on the other; and (c) three-factor, where the items from each scale loaded on their respective factor. Although the fit indices for all three models were acceptable, the three-factor model had the best fit indices. Consequently, this three-factor model was tested with sample 4 and the results of the CFA also supported the internal structure in which it obtained acceptable fit indices. These results support the internal structure of the developed scales based on the motivational phase of the IMV model of suicidal behavior (O'Connor, 2011), which implicitly considers suicidal ideation, feelings of defeat and entrapment as unique and independent, but related, constructs.

The present study provides insight on measurement invariance of the three brief scales across gender, age, job position, type of organization, and type of employment. We tested the measurement invariance of the suicidal behavior brief scales among employees of different organizations in Puerto Rico. Exploration on the first two levels revealed configural and metric invariance (i.e., weak measurement invariance) and scalar invariance (i.e., strong measurement invariance) of the three-factor model across gender, age, job position, and type of organization. Metric invariance is important to ensure the measure across multiple groups is on the same scale, or the factors are measured in the same way in all groups (e.g., Wang & Wang, 2012). Scalar invariance refers to the item intercepted being invariant across multiple groups in the present study. This indicates that none of the groups tends to be systematically higher or lower on the items of scales than other groups (Wang & Wang, 2012). The present study met both invariance requirements. These results confirm that the compared groups had an equivalent understanding on each of the scale's items, which is an important prerequisite for making a meaningful comparison between groups on these suicide behaviors at work. Researchers have argued that error variance invariance (i.e., strict measurement invariance) is not required for substantive analyses in many disciplines and such invariance is considered unnecessary (Wang & Wang, 2012).

To establish convergent and divergent validity of the three suicidal behaviors at work brief scales, first, we calculated the AVE, MSV, and ASV. The AVE ≥ .50 indicates that the items share a high proportion of the variance, the higher the value of the AVE, the lower the error variance (Fornell & Larcker, 1981). Therefore, the indicators of each scale developed share a high proportion of variance supporting their convergent validity. Moreover, the AVE's value of the three scales were greater than the MSV and ASV values, supporting the divergent validity of the scale as some authors suggest (e.g., Fornell & Bookstein, 1982; Fornell & Larcker, 1981). In addition, observed correlation directions between the three suicidal behaviors at work brief scales with rumination, depression, anxiety, and social desirability were as hypothesized. Current results shown that defeat and entrapment are related to suicidal ideation as some of the literature has found (e.g., Rosario-Hernández et al., 2019; Taylor, Wood et al., 2010). In the case of the defeat and entrapment constructs, some authors have conceptualized them as one-factor construct (e.g., Taylor et al., 2010) because of their high correlation. Therefore, the large correlation between them was expected in the present study. In fact, results show that defeat and entrapment appear to be two constructs that are closely related (e.g., Taylor et al., 2009), while still having distinct qualities that set them apart. WRSI, defeat, and entrapment are theoretically related within them, as shown by our findings and certain literature (e.g., O'Connor et al., 2016). Moreover, the relationship found in the present study between the results of the three short self-report scales and other constructs tend to support the convergent validity of these scales, as some literature suggest for depression (e.g., Tang et al., 2010), anxiety (e.g., Tang et al., 2010), and rumination (e.g., Treynor et al., 2003), even when some literature argue that this relationship is mediated by feelings of entrapment (Teismann & Forkmann, 2017). On the other hand, the relationship between the social desirability and the suicidal behaviors at work scales was negative, but with much lower correlation coefficients when compared to other studies' results (e.g., Caputo, 2017; Curns, 2014). Nevertheless, these results support the divergent validity of the three new developed scales.

Regarding reliability, the coefficients alpha and omega, the levels obtained can be considered as excellent from a general perspective and considering the interaction between the small number of items, especially the Defeat and Entrapment Scales, the sample size and the values obtained (Ponterotto & Ruckdeschel, 2007). These scales' primary usage is for group applications, but because their coefficients are high (i.e., ≥.85), it may be assumed that the likelihood of error is low, even in cases when judgements on individual subjects are required (Ponterotto & Ruckdeschel, 2007). However, given the similarity of ∝ and ω, it is considered that any differences in the factor loadings were minor and did not significantly affect how close one coefficient was to the other (Hayes & Coutts, 2020). This distance is typically related to the level of factorial item loading equality, or tau-equivalence, which is a prerequisite for validating ∝ coefficient (Hayes & Coutts, 2020). The calculation of internal consistency can be done successfully using ∝ and without the need for SEM modeling or SEM modeling methodologies to estimate ω, according to an implication of this similarity. This application can be induced to other contexts if the prerequisites for application in future usage and the data cleaning are successful.

Theoretical and Practical Implications

This study makes a valuable contribution to current research on suicidal behavior at work by developing and validating three robust scales to measure WRSI, feelings of defeat and entrapment. Unlike previous measurement scales of suicidal behavior, the scale developed in this study, especially the WRSI scale, is more appropriate for studying suicidal ideation related to work because it incorporates causal attributions to work. Thus, in comparison to other suicidal ideation measures, the items selected for the WRSI scale explicitly ask respondents whether they attribute their suicidal ideation to wok; therefore, the WRSI scale has a protocol to help dismiss suicidal ideation attributed to nonwork sources (e.g., a conflictual spousal relationship) or a source the respondent cannot identify. Also, these scales might contribute to the study of suicidal behavior at work in the prevention and control in the foreseeable future by providing brief, but robust measures of these constructs. In addition, the developed and validated scales include three important constructs of the motivational phase of the IMV model of suicidal behavior from which scores derived and they appear to have excellent reliability and evidence of their validity based on their items. The results also indicate that these three constructs are essential for the measurement of the motivational phase of the IMV model of suicidal behavior, and the validated suicidal behavior at work scales derived from this study can be used as a primary benchmark tool to help in the study of suicidal behavior at work to develop suicide prevention programs in workplaces. Thus, it would contribute to mitigate the risk of suicide and the overall well-being of employees using, at least in part, the IMV model of suicidal behavior. The implementation of the suicidal ideation at work scales can also provide rich feedback to policymakers, mental health professionals, and managers to plan interventions about suicidal behavior at work. In terms of prevention, having valid and reliable tools to identify the risk of suicide is desirable (Vecco at al., 2021) and these developed suicidal behavior at work scales have the potential to help in this end.

Limitations and Recommendations

When evaluating the findings, it is important to consider the current study's numerous flaws. First, because the sample was not chosen randomly and the population resemblance was not confirmed, the population representativeness cannot be assured.Therefore, it is important to cross validate these results with other samples of Puerto Rican employees. Second, because multiple procedures can create varying percentages of type I and type II errors, it may be necessary to investigate how other approaches compare to the single procedure used to examine measurement invariance (i.e., differential operation approach of items). Finally, the reliability evaluation of the stability of the scores was not completed. Consequently, to finish the evaluation of this element, the score's repeatability over time using a test-retest methodology should be investigated.

Conclusion

The final version of the suicidal behavior at work scales consists of three brief measures of work-related suicidal ideation, feelings of defeat and entrapment that are essential constructs of the motivational phase of the IMV of suicidal behavior. The scales' reliability, the evidence of their validity, and the strong measurement invariance between groups (i.e., gender, age, job position, type of organization, and type of employment) suggest that these measures are robust to be used in the occupational health psychology field in the context of organizations in Puerto Rico.

References

Avendaño-Prieto, B. L., Pérez-Prada, M., Vianchá-Pinzón, M., Martínez-Baquero, L., & Toro, R. (2018). Propiedades psicométricas del inventario de ideación suicida positiva y negativa PANSI. *Revista Evaluar, 18*(1). https://doi.org/10.35670/1667-4545.v18.n1.19767

Browne, M. W., & Cudeck, R. (1993). Alternative ways of assessing model fit. In K. A. Bollen & J. S. Long (Eds.), *Testing structural equation models* (pp. 136-162). SAGE.

Byrne, B. M. (2016). *Structural equation modeling with AMOS: Basic concepts, applications, and programming* (3rd ed.). Routledge. https://doi.org/10.4324/9781315757421

Caputo, A. (2017). Social desirability bias in self-reported well-being measures: Evidence from an online survey. *Universitas Psychologica, 16*(2). https://doi.org/10.11144/Javeriana.upsy16-2.sdsw

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

Cohen, J. (1988). *Statistical power analysis for the behav-*

ioral sciences (2nd ed.). Routledge Academic. https://doi.org/10.4324/9780203771587

Commission for the Prevention of Suicide. (2016). *Estadísticas preliminaries de casos de suicidio Puerto Rico, enero – diciembre 2016*. Estado Libre Asociado de Puerto Rico, Departamento de Salud. https://www.dropbox.com/s/va6cwyh84ajamu6/Estadisticas%20Suicidios%20en%20Puerto%20Rico%20Diciembre%202016.pdf?dl=0

Cropley, M., Michalianou, G., Pravettoni, G., & Millward, L. J. (2012). The relation of post-work ruminative thinking with eating behaviour. *Stress & Health, 28*(1), 23-30. https://doi.org/10.1002/smi.1397

Curns, D. B. (2014). *A validity study of the Reasons for Life Scale with emerging adult college students*. Unpublished dissertation for University of Alaska Fairbanks and Anchorage. https://scholarworks.alaska.edu

DeVellis, R. F. (2017). *Scale development: Theory and applications* (4th ed.). SAGE.

Fabrigar, L. R., Wegener, D. T., MacCallum, R. C., & Strahan, E. J. (1999). Evaluating the Use of Exploratory Factor Analysis in Psychological Research. *Psychological Methods, 4*, 272-299. https://doi.org/10.1037/1082-989X.4.3.272

Fornell, C., & Bookstein, F. L. (1982). Two structural equation models: LISREL and PLS applied to consumer exit-voice theory. *Journal of Marketing Research, 19*(4), 440-452. https://doi.org/10.2307/3151718

Fornell, C., & Larcker, D. F. (1981). Evaluating structural equation models with unobservable variables and measurement error. *Journal of Marketing Research, 18*(1), 39-50. https://doi.org/10.2307/3151312

Hair, J. F., Black, W. C., Babin, B. J., & Anderson, R. E. (2019). *Multivariate data analysis* (8th ed.). Cengage Learning EMEA.

Hayes, A. F., & Coutts, J. J. (2020). Use Omega rather than Cronbach's Alpha for estimating reliability. But… *Communication Methods and Measures, 14*(1), 1-24. https://doi.org/10.1080/19312458.2020.1718629

Hu, L.-T., & Bentler, P. M. (1995). Evaluating model fit. In R. H. Hoyle (Ed.), *Structural equation modeling: Concepts, issues, and applications* (pp. 76-99). SAGE.

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

IBM Corp. (2021). IBM SPSS Statistics for Windows, Version 28.0. Armonk, NY: IBM Corp.

Kasl, S. V., & Jones, B. A. (2003). An epidemiological perspective on research design, measurement, and surveillance strategies. In J. Campbell Quick & L. E. Tetrick (Eds.), *Handbook of Occupational Health Psychology* (pp. 377-398). American Psychological Association.

Kline, P. (1994). *An easy guide to factor analysis* (1st ed.). Routledge.

Kline, R. B. (2016). *Principles and practice of structural equation modelling* (4th ed.). The Guilford Press.

Klonsky, E. D., Saffer, B. Y., & Bryan, C. J. (2018). Ideation-to-action theories of suicide: A conceptual and empirical update. *Current Opinion in Psychology, 22*, 38-43. https://doi.org/10.1016/j.copsyc.2017.07.020

Kroenke, K., Spitzer, R. L., & Williams, J. B. W. (2001). The PHQ-9: Validity of a brief depression severity measure. *Journal of General Internal Medicine, 16*(9), 606-613. https://doi.org/10.1046/j.1525-1497.2001.016009606.x

Li, C.-H. (2016a). Confirmatory factor analysis with ordinal data: Comparing robust maximum likelihood and diagonally weighted least squares. *Behavior Research Methods, 48*(3), 936-949. https://doi.org/10.3758/s13428-015-0619-7

Li, C.-H. (2016b). The performance of ML, DWLS, and ULS estimation with robust corrections in structural equation models with ordinal variables. *Psychological Methods, 21*(3), 369-387. https://doi.org/10.1037/met0000093

Littlewood-Zimmerman, H. F., & Bernal-García, E. R. (2011). *Mi primer modelamiento de ecuación* *estructural: LISREL*. Centro de Investigación en Comportamiento Organizacional (CINCEL).

MacCallum, R. C., Browne, M. W., & Sugawara, H. M. (1996). Power analysis and determination of sample size for covariance structure modeling. *Psychological Methods, 1*(2), 130-149. https://psycnet.apa.org/doi/10.1037/1082-989X.1.2.130

Marsh, H. W., Balla, J. R., & Hau, K. T. (1996). An evaluation of incremental fit indexes: A clarification of mathematical and empirical properties. In G. A. Marcoulides & R. E. Schumacker (Eds.), *Advanced structural equation modeling techniques* (pp. 315-353). Lawrence Erlbaum.

Mortali, M. G., & Moutier, C. (2019). Suicide prevention in the workplace. In M. B. Riba, S. V. Parikh & J. F. Greden (Eds.), *Mental health in the workplace: Strategies and tools to optimize outcomes*. Springer.

Muthén, L. K., & Muthén, B. O. (2012). *Mplus User's Guide: Statistical Analysis with Latent Variables* (7th ed.). Muthén & Muthén.

O'Connor, R. C. (2011). Towards an integrated motivational-volitional of suicidal behavior. In R. C. O'Connor, S. Platt & J. Gordon (Eds.), *International handbook of suicide prevention: Research, policy and practice* (pp. 181-198). Wiley-Blackwell.

O'Connor, R. C., Cleare, S., Eschle, S., Wetherall, K., & Kirtley, O. J. (2016). The integrated motivational-volitional model of suicidal behavior: Update. In R. C. O'Connor & J. Pirkis (Eds.), *The International Handbook of Suicide Prevention* (pp. 220-240). John Wiley & Sons.

O'Connor, R. C., & Nock, M. K. (2014). The psychology of suicidal behaviour. *The Lancet Psychiatry, 1*(1), 73-85. https://doi.org/10.1016/S2215-0366(14)70222-6

Otsuka, Y., Nakata, A., Sakurai, K., & Kawahito, J. (2016). Association of suicidal ideation with job demands and job resources: A large cross-sectional study of Japanese workers. *International Journal of Behavioral Medicine, 23*(4), 418-426. https://doi.org/10.1007/s12529-016-9534-2

Peterson, C., Stone, D. M., Marsh, S. M., Schumacher, P. K., Tiesman, H. M., McIntosh, W. L.-K., Lokey, C. N., Trudeau, A.-R.T., Bartholow, B., & Luo, F. (2018). Suicides rate by major occupational group – 17 states, 2012 and 2015. *Morbidity and Mortality Weekly Report, 67*(45), 1253-1260. http://dx.doi.org/10.15585/mmwr.mm6745a1

Ponterotto, J. G., & Ruckdeschel, D. E. (2007). An overview of Coefficient Alpha and a reliability matrix for estimating adequacy of internal consistency coefficients with psychological research measures. *Perceptual and Motor Skills, 105*(3), 997-1014. https://doi.org/10.2466/pms.105.3.997-1014

Rigdon, E. E. (1995). A necessary and sufficient identification rule for structural models estimated in practice. *Multivariate Behavioral Research, 30*(3), 359-383. https://doi.org/10.1207/s15327906mbr3003_4

Rosario-Hernández, E., & Rovira-Millán, L. V. (2002). Desarrollo y validación de una escala para medir actitudes hacia el retiro. *Revista Puertorriqueña de Psicología, 13*, 45-60. https://www.repsasppr.net/index.php/reps/index

Rosario-Hernández, E., Rovira-Millán, L. V., & Merino-Soto, C. (2021). Review of the internal structure, psychometric properties, and measurement invariance of the Work-Related Rumination Scale – Spanish Version. *Frontiers in Psychology, 12*. https://doi.org/10.3389/fpsyg.2021.774472

Rosario-Hernández, E., Rovira-Millán, L. V., Vega-Vélez, S., Zeno-Santi, R., Farinacci-García, P., Centeno-Quintana, L., Navedo-Santos, J., Feliciano-Toro, B. P., De Jesús-Caraballo, R., Morell-Fausto, J., Cepeda-Fax, S., Cintrón-Lugo, M., Colón-Burgos, L., Sánchez-Ortiz, N., Sánchez-Collazo, I., Díaz-Pla, L., Toledo-Medina, M. A., Flores-Quirós, A. S., & Pagán-Torres, O. M. (2019). Exposure to workplace bullying and suicidal ideation: An exploratory study. *Journal of Applied Structural Equation Modeling, 3*(1), 55-75. https://doi.org/10.47263/JASEM.3(1)06

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

Spector, P. E. (1992). *Summated rating scale construction: An introduction*. SAGE.

Spitzer, R. L., Kroenke, K., Williams, J. B. W., & Löwe, B. (2006). A brief measure for assessing generalized anxiety disorder. Tthe GAD-7. *Archives of Internal Medicine, 166*(10), 1092-1097. https://doi.org/10.1001/archinte.166.10.1092

Stark, S., Chernyshenko, O. S., & Drasgow, F. (2006). Detecting differential item functioning with confirmatory factor analysis and item response theory: Toward a unified strategy. *Journal of Applied Psychology, 91*(6), 1292-1306. https://doi.org/10.1037/0021-9010.91.6.1292

Suominen, K., Isometsä, E., Ostamo, A., & Lönnqvist, J. (2004). Level of suicidal intent predicts overall mortality and suicide after attempted suicide: A 12-year follow-up study. *BMC Psychiatry, 4*, Article 11. https://doi.org/10.1186/1471-244X-4-11

Tang, N. K. Y., Goodchild, C. E., Hester, J., & Salkovskis, P. M. (2010). Mental defeat is linked to interference, distress and disability in chronic pain. *Pain, 149*(3), 547-554. https://doi.org/10.1016/j.pain.2010.03.028

Taylor, P. J., Gooding, P. A., Wood, A. M., Johnson, J., Pratt, D., & Tarrier, N. (2010). Defeat and entrapment in schizophrenia: The relationship with suicidal ideation and positive psychotic symptoms. *Psychiatry Research, 178*(2), 244-248. https://doi.org/10.1016/j.psychres.2009.10.015

Taylor, P. J., Wood, A. M., Gooding, P., & Tarrier, N. (2010). Appraisals and suicidality: The mediating role of defeat and entrapment. *Archives of Suicide Research, 14*(3), 236-247. https://doi.org/10.1080/13811118.2010.494138

Taylor, P. J., Wood, A. M., Gooding, P., Johnson, J., & Tarrier, N. (2009). Are defeat and entrapment best defined as a single construct? *Personality and Individual Differences, 47*(7), 795-797. https://doi.org/10.1016/j.paid.2009.06.011

Teismann, T., & Forkmann, T. (2017). Rumination, entrapment and suicide ideation: A mediational model. *Clinical Psychology & Psychotherapy, 24*(1), 226-234. https://doi.org/10.1002/cpp.1999

Tiesman, H. M., Konda, S., Hartley, D., Chaumont-Menéndez, C., Ridenour, M., & Hendricks, S. (2015). Suicide in U.S. workplaces, 2003-2010: A comparison with non-workplace suicides. *American Journal of Preventive Medicine, 48*(6), 674-682. https://doi.org/10.1016/j.amepre.2014.12.011

Treynor, W., Gonzalez, R., & Nolen-Hoeksema, S. (2003). Rumination reconsidered: A psychometric analysis. *Cognitive Therapy and Research, 27*(3), 247-259. https://doi.org/10.1023/A:1023910315561

Tucker, L. R., & Lewis, C. (1973). A reliability coefficient for maximum likelihood factor analysis. *Psychometrika, 38*(1), 1-10. https://doi.org/10.1007/BF02291170

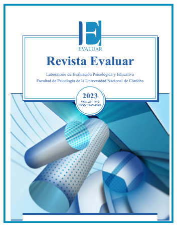
Vecco, G. A., Flores-Kanter, P. E., & Luque, L. E. (2021). Análisis psicométrico del Inventario de Orientación Suicida ISO-19, en adolescentes cordobeses escolarizados. *Revista Evaluar, 21*(1), 39-52. https://doi.org/10.35670/1667-4545.v21.n1.32831

Wang, J., & Wang, X. (2012). *Structural equation modeling: Applications using Mplus* (1st ed.). John Wiley & Sons.

World Health Organization. (2021). *Suicides worldwide in 2019: Global health estimates*. https://www.who.int/publications

2023, Vol. 23, No. 2
ISSN 1667-4545
Recuperado de https://revistas.unc.edu.ar/index.php/revaluar

Laboratorio de Evaluación Psicológica y Educativa
Facultad de Psicología - Universidad Nacional de Córdoba



Psychometric Properties of the Psychological Capital Questionnaire (PCQ-12) in Dominican Secondary School Students

Propiedades psicométricas del Cuestionario de Capital Psicológico (PCQ-12) en estudiantes de secundaria dominicanos

Introduction
Method
Results
Discussion
References

Sara Martínez-Gregorio * [1] , Diana Gómez-Cruz [1], Amparo Oliver [1]

1 - Departamento de Metodología de las Ciencias del Comportamiento, Universitat de València, España.

Recibido: 25/02/2023 **Revisado:** 10/04/2023 **Aceptado:** 17/04/2023

Abstract

This research aimed to analyze the psychometric properties of the 12-item Psychological Capital Questionnaire (PCQ-12) in secondary school students from the Dominican Republic. The questionnaire was completed by a total of 708 students aged 11 to 19 (M = 15.49 years; SD = 1.58), with 64.7% being females. Through Confirmatory Factor Analysis (CFAs), the different dimensionalities proposed in the previous literature were tested and the structure of four factors with a second-order factor was retained. Next, the reliability of the dimensions was studied and problems in optimism were identified, especially, in resilience. The second-order structure showed to be invariant to the students' gender, supporting its absence of gender bias. Consequently, the present study supports the use of the scale to measure the Psychological Capital as a second-order construct, but calls for the development of research that improves the measuring of resilience.

Keywords: *psychological capital, adolescents, resilience, optimism, self-efficacy, hope, psychometric properties, invariance*

Resumen

El objetivo de esta investigación fue analizar las propiedades psicométricas del Cuestionario de Capital Psicológico PCQ-12 en estudiantes de educación secundaria de República Dominicana. El cuestionario fue completado por un total de 708 estudiantes de entre 11 y 19 años (M = 15.49 años; DT = 1.58) entre los cuales el 64.7% fueron mujeres. Mediante Análisis Factoriales Confirmatorios (AFCs), se pusieron a prueba las distintas dimensionalidades propuestas en la literatura previa y se retuvo la estructura de cuatro factores con un factor de segundo orden. A continuación, se estudió la fiabilidad de las dimensiones y se identificaron problemas en las dimensiones de optimismo y, especialmente, resiliencia. La estructura de segundo orden mostró ser invariante al género de los estudiantes, lo que respaldó su ausencia de sesgo de género. Consecuentemente, se respalda el uso de la escala para la medición del Capital Psicológico como constructo de segundo orden y se invita al desarrollo de investigaciones que mejoren la medición de la resiliencia.

Palabras clave: *capital psicológico, adolescentes, resiliencia, optimismo, autoeficacia, esperanza, propiedades psicométricas, invarianza*

***Correspondence to:** Sara Martínez-Gregorio, sara.martinez@uv.es, Departamento de Metodología de las Ciencias del Comportamiento, Facultad de Psicología y Logopedia, Av. Blasco Ibáñez, 21, 46010, Valencia (Spain). Teléfono: +34 963864468.

Author's note: Las autoras no tienen conflictos de interés.

How to cite: Martínez-Gregorio,S., Gómez-Cruz, D., & Oliver, A. (2023). Psychrometric Properties of the Psychological Capital Questionnaire (PCQ-12) in Dominican Secondary School Students. *Revista Evaluar*, *23*(2), 19-27. Retrieved from https://revistas.unc.edu.ar/index.php/revaluar

Acknowledgments: Sara Martínez-Gregorio es beneficiaria del Programa de Formación del Profesorado Universitario del Ministerio de Universidades (FPU18/03710).

Participaron en la edición de este artículo: Vanesa Mariela Toledo, Andrea Suárez, Eugenia Barrionuevo, Abigail Peréz, Gloria Rebeca Nieve, Florencia Ruiz, Jorge Bruera.

Introduction

Strengths such as self-efficacy, optimism, hope, and resilience are essential when it comes to positively assessing a circumstance or to predicting the success of individuals based on aspects such as perseverance and effort (Azanza et al., 2014). In this context, Psychological Capital arises. It is a construct that alludes to a state of positive individual development shaped by four dimensions, which correspond precisely to characteristics such as self-efficacy, optimism, hope, and resilience (Luthans et al., 2007; Luthans et al., 2015).

To understand the delimitation of the construct, it is worth specifying the definitions of each of its four dimensions to deepen and have a closer view of the elements that constitute Psychological Capital (Luthans et al., 2007; Stajkovic & Luthans, 2003). We understand self-efficacy as the person's confidence and effort to face a challenging task successfully (Bandura, 1997). Optimism consists of making positive attributions about current and future success. On the other hand, hope consists of people's perseverance toward their goals and redirecting alternatives to achieve them successfully (Vuyk & Codas, 2019). Finally, resilience is defined as the ability to sustain oneself and cope with problems and adversities that arise in individuals' lives (Peña-Contreras et al., 2020).

The construct of Psychological Capital emerged in the context of organizations (Luthans & Youssef, 2004) and many studies and research have been conducted on its impact, having identified a positive relationship between this concept and psychosocial and organizational variables such as leadership, confidence, creativity, and performance, among others (Clapp-Smith et al., 2009; Rego et al., 2012). However, although some authors have indeed transferred the application of this idea to the student population, research in the educational context is recent and limited compared to that developed in the organizational literature (Martínez et al., 2021; Schönfeld & Mesurado, 2020; Tomás et al., 2022). Previous research shows that there is a significant relationship between Psychological Capital and variables such as students' grade point average, their satisfaction with school, their development, retention, and success, and even with academic performance (Azanza et al., 2014; Carmona-Halty et al., 2019; Datu et al., 2018; Luthans et al., 2012).

In this context of growing academic interest in the construct and, specifically, in adolescents, it is essential to have an adequate conceptualization and measurement of the Psychological Capital through psychometric instruments adapted and validated for this new use. Among the scales present in the literature for measuring Psychological Capital, the 12-item Psychological Capital Questionnaire (PCQ-12; Avey et al., 2011) is currently one of the most widely used questionnaires to measure Psychological Capital in adolescents. Despite the popularity of the PCQ-12, there is some debate about its structure and reliability problems (Djourova et al., 2019). There are recent studies about the psychometric properties of this instrument in academic contexts (Martínez et al., 2021; Schönfeld & Mesurado, 2020; Tomás et al., 2022). Martínez et al. (2021) found problems concerning the reliability of the scale in two of its dimensions and factor loadings when testing the model with a second-order factor and not having it compared with the alternative four correlated factors. In contrast, Schönfeld and Mesurado (2020) found no reliability problems, although they did not test the structure of the four correlated factors.

Along the same lines, Tomás et al. (2022) compared the three competitive models around which there has been controversy: one factor, four correlated factors, and a second-order structure through estimations with Bayesian methods. Additionally, these authors introduced in the literature the possibility of a bifactor model. Conclusions of the research indicated that the second-order structure is the one most supported by the evidence, showing a similar fit to the bifactor but with a more parsimonious structure. In addition, the scale also found difficulties in measuring resilience in adolescents by showing low reliability scores.

To date, these questionnaires have not been used in the Dominican Republic context where the consideration of Psychological Capital has been scarce in national studies. There are no studies that introduced some of its dimensions, but none considered the entire construct. For example, Tomás et al. (2020) showed the relevance of hope and self-efficacy in the academic context, both being precursors of commitment and, indirectly, of self-concept and academic performance. These previous studies serve as an encouragement to show the potential of conceptualizing the positive student state in a more complex way, including optimism and resilience. The lack of consideration of Psychological Capital means that, to date, there are no psychometric studies conducted in the Dominican context that report on the suitability of the scale for use with students.

Additionally, these psychometric studies should examine the absence of gender bias in the measurement of Psychological Capital, as Avey (2014) suggested, by considering gender differences when investigating Psychological Capital in adolescents. These gender differences can only be studied if the scale works comparably for both genders, an issue that has not been tested yet in previous psychometric studies with adolescents.

Consequently, this research work proposes the study of the psychometric properties of the PCQ-12 in a sample of adolescents from the Dominican Republic. For this purpose, (1) the descriptive statistics of the items were calculated; (2) the dimensionality of the scale was explored; (3) the reliability of its dimensions was studied; and (4) a routine was established to evaluate the gender invariance of the scale.

Method

Participants and procedure

The study sample consisted of 708 secondary school students from the Dominican Republic. The mean age was 15.49 years (SD = 1.58), with a minimum of 11 and a maximum of 19 years. From the sample, 64.7% were female (n = 458) and 34% male (n = 241), and a total of 9 students did not declare their gender (1.3%).

Instruments

For this research, the instrument used was the 12-item Psychological Capital Questionnaire (PCQ-12; Avey et al., 2011) in its adaptation for Spanish-speaking secondary school students (Tomás et al., 2022). The questionnaire presents a total of 12 items through which the four dimensions of psychological capital are assessed: self-efficacy (items 1, 2, and 3), hope (items 4, 5, 6, and 7), resilience (items 8, 9, and 10), and optimism (items 11 and 12). The adaptation for secondary school students differs from the original for adults by replacing references to work with "studies". The response format is a five-anchor Likert scale ranging from *Strongly disagree* to *Strongly agree*.

Along with the instrument, a series of sociodemographic data, such as the gender and age of the participants, were collected.

Data analysis

The study of the psychometric properties in the questionnaire was conducted by following

Table 1
Descriptive statistics of the items.

	M	SD	g^1	g^2	1	2	3	r_{it}
SE1	3.93	0.95	-0.98	0.89				.65
SE2	3.95	0.95	-0.96	0.71	.65			.76
SE3	3.94	0.92	-0.95	0.84	.54	.68		.67
HO1	4.06	0.80	-1.20	2.51				.45
HO2	3.79	0.89	-0.45	0.03	.33			.54
HO3	4.20	0.81	-1.46	3.52	.42	.41		.57
HO4	3.85	0.92	-0.95	0.96	.33	.51	.48	.58
RE1	3.76	0.93	-0.76	0.44				.21
RE2	3.44	1.14	-0.49	-0.57	.17			.28
RE3	3.68	0.96	-0.79	0.49	.16	.26		.28
OP1	3.80	0.92	-0.71	0.42				.49
OP2	3.89	0.87	-0.80	0.97	.49			.49

Note. M = Mean; SD = Standard deviation; g^1 = Kurtosis; g^2 = Skewness; r_{it} = Item-total correlation. All correlations in the table were statistically significant $p < .001$.

a series of steps. First, the descriptive statistics of the items that constitute the scale were calculated (mean, standard deviation, skewness, kurtosis, inter-item correlations, and corrected item-total correlation). Next, its factorial structure was studied. For this purpose, a series of Confirmatory Factor Analyses (CFA) were tested with the different factorial solutions observed in the previous literature: (a) one factor, (b) four correlated factors, and (c) four first-order factors with a second-order factor.

The estimation method used was Maximum Likelihood Robust (MLR). The adequacy of the AFCs was evaluated by considering several fit indices: chi-square, CFI, RMSEA, and SRMR. A CFI is considered adequate when it presents a value above .90, with a value above .95 being desirable. For the RMSEA and SRMR, adequate values should be below .08 (Marsh et al., 2004). Once the most appropriate factor structure was identified, the reliability of the dimensions was studied. Composite Reliability Indexes (CRI) were calculated to estimate reliability for each dimension and the total scale. The formula presented by Raykov and Marcoulides (2012) was used to calculate the reliability of the second-order factor.

To conclude, the invariance of the scale by gender was tested as a method for studying the possible differential functioning of the items. For this purpose, since it is a second-order structure, the procedure proposed by Chen et al. (2005) and the syntax presented by Dimitrov (2010) were followed. These authors propose five nested models for the study of invariance. First, it tests the configural invariance of the scale, thus checking the fit of the structure for both genders. Next, it tests the metric invariance of the factor loadings of the items (Metric 1). If the metric invariance at the item level is satisfied, it continues to fix factor loadings of the first-order factors (Metric 2). If met, the scalar invariance of the items is tested, setting their intercepts equal for both groups

Table 2
Fit indices of the models proposed.

Model	χ^2	*df*	*p*	CFI	RMSEA	90% CI	SRMR
A factor	317.676	54	< .01	.86	.08	.07, .09	.05
Four-correlated factors	110.675	48	< .01	.97	.04	.03, .05	.03
Second-order model	114.647	50	< .01	.97	.04	.03, .05	.03

(Scalar 1). Finally, the intercepts of the first-order factors are additionally fixed (Scalar 2). The different nested models are compared using two complementary procedures (Little, 1997). On the one hand, a formal statistical test is performed using chi-square differences, the absence of statistical significance being the evidence of invariance. This method has been criticized in the literature for being too strict, identifying trivial differences in practice (Cheung & Rensvold, 2002). Therefore, an assessment of the changes in the model fit indices was performed. To consider that the invariance assumption is met, the CFI should not vary by more than .01 (Wang & Wang, 2012).

Descriptive analyses were performed with the statistical program IBM SPSS Statistics version 28, while CFAs were performed in the program Mplus 8.6 (Muthén & Muthén, 2017).

Results

Descriptive statistics

Table 1 shows the descriptive statistics of all the items included in the scale, as well as the correlations between the items that compose each dimension and the item-factor correlations for each of them. As it can be observed, the mean scores of the subjects on the items are above the centre point (3) in all cases. All the items show positive, statistically significant, with higher correlations with the rest of the items of their dimension, except for those that make up the resilience dimension. Despite being positive and statistically significant, the correlations of the dimension items are below .3.

Factor structure

Table 2 shows the fit results of the different models proposed. The single-factor model did not show an adequate fit to the data, its CFI being below .9. In contrast, both the four-factor correlated model and the model with a second-order factor showed a good fit to the data, with CFIs above .95 and the RMSEA and SRMR below .05. Given their fit equality, the second-order model is chosen as the best model because of its parsimony.

Figure 1 shows the results for the four-factor correlated model and the second-order model. In both cases, all item factor loadings in their corresponding factor are above .3, ranging from .37 (for item 3 of the resilience dimension) to .86 (of the second item of the self-efficacy dimension). All of them were statistically significant ($p < .001$). In the case of the second-order model, all the loadings of the first-order factors are high, the lowest being that of the self-efficacy dimension.

Reliability

Reliability was calculated using the CRI for each dimension and the total psychological capital. The self-efficacy and hope dimensions showed adequate scores above .7 (.84 and .74,

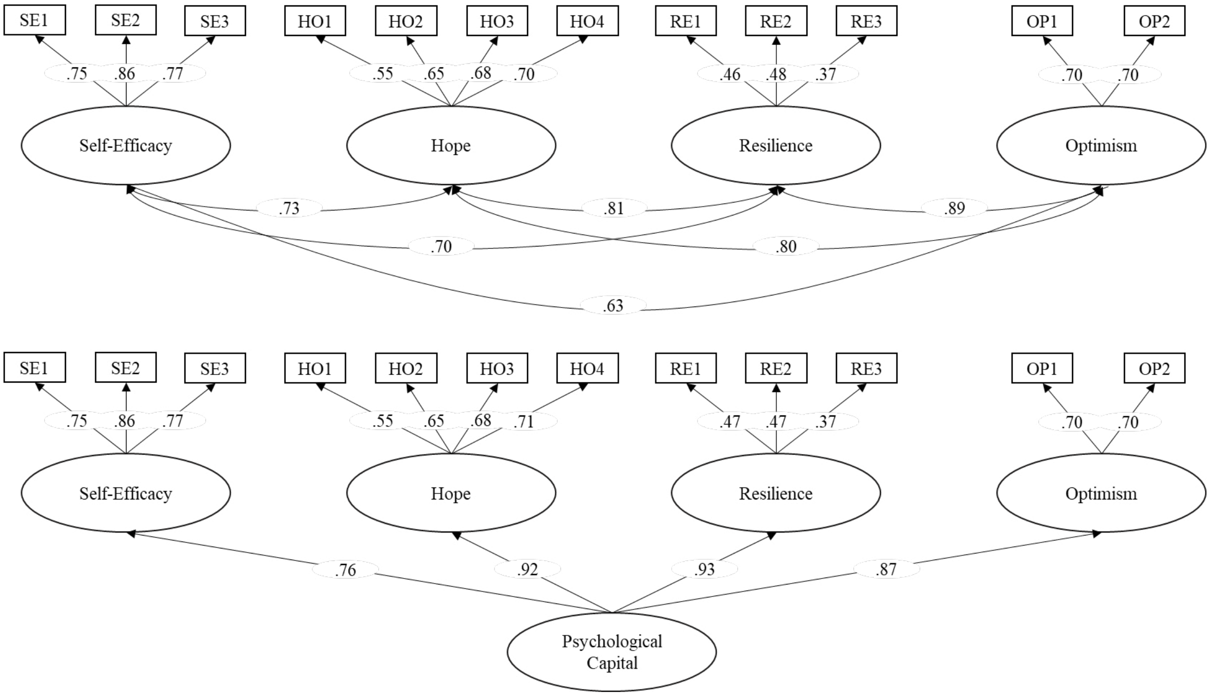


Figure 1
Four-factor correlated and second-order factor models.

respectively). In contrast, the results for the resilience (.42) and optimism (.66) dimensions were lower, with the resilience result being particularly poor. The reliability score for the psychological capital factor is .69, very close to .7.

Gender invariance

Once the second-order factor model was established as the most parsimonious among those with the best fit, its invariance across gender was tested. The fit of the different nested models is shown in Table 3. As it can be observed, the model is invariant for boys and girls at the configural and metric levels. That is, the structure is adequate for the data of both groups and the factor loadings are identical. When we reach the scalar invariance of the item intercepts (Scalar 1), we see that the model slightly worsens the fit, the CFI dropping above .01, although it is very close. Consequently, we continue with the last step, where no differences were found between the intercepts of the first-order factors in both groups.

Discussion

Psychological Capital is a construct that has recently burst into research related to adolescent academic success (Azanza et al., 2014; Carmona-Halty et al., 2019; Datu et al., 2018; Luthans et al., 2012). This growing interest has been accompanied by the development of studies for improving its measurement. Specifically, in recent years, different researchers have focused on the psychometric properties of one of the most widely used instruments in the literature of the PCQ-12. Its dimensionality and reliability have been tested in samples of students from Spain, Chile, and Argentina (Martínez et al., 2021; Schönfeld

Table 3
Goodness-of-fit indices for each group (men and women) and a set of nested models to test gender invariance.

Model	χ^2	*df*	*p*	$\Delta\chi^2$	Δgl	*p*	CFI	ΔCFI	SRMR	ΔSRMR	RMSEA	ΔRMSEA	90% CI
Configural	186.850	100	< .001	--	--	--	.955	--	.042	--	.050	--	.039-.061
Metric 1	195.759	108	< .001	9.441	8	.31	.955	.000	.052	.010	.048	-.002	.037-.059
Metric 2	197.252	111	< .001	2.080	3	.55	.956	.001	.054	.002	.047	-.001	.036-.058
Scalar 1	230.788	122	< .001	36.778	11	< .001	.944	-.012	.062	.008	.051	.004	.040-.060
Scalar 2	231.101	123	< .001	.032	1	.86	.944	.000	.062	.000	.050	-.001	.040-.060

Note. *df* = degrees of freedom; Δ = differences.

& Mesurado, 2020; Tomás et al., 2022). These studies highlighted some controversies regarding its factorial structure and some limitations with the reliability of some of its dimensions. Two aspects require more attention: (1) the Psychological Capital construct has received less attention in the Caribbean context and, specifically, there are no psychometric studies for this population, and (2) the gender invariance of the scale has never been tested in adolescents. Given this situation, this study was developed.

Regarding the factorial structure, the alternatives of four correlated factors and that of a second-order factor presented an identical fit. Consequently, the second-order model is considered more appropriate as it is more parsimonious and supports the use of Psychological Capital as a unitary construct that has been proposed in previous research (Carmona-Halty et al., 2019; Slåtten et al., 2021). This result is consistent with that shown in the Spanish and Argentinean samples (Schönfeld & Mesurado, 2020; Tomás et al., 2022).

Regarding reliability, previous authors identified repeated reliability problems in the Dominican sample. Specifically, the dimensions of resilience and optimism did not show an adequate reliability score. The result for optimism is not alarming, being close to .7, but the resilience dimension shows poor reliability. These results replicate those found by Tomás et al. (2022). As in this study, it is item 10 that presents the greatest problems. This study supports the generalization of the problem identified by Tomás et al. (2022) in the Spanish-speaking context. The content of item 10 could be confusing or could simultaneously pose two independent questions that prevent the student from answering adequately.

Despite the problems identified, the scale performs equally well in boys and girls. The invariance routine shows that the scale structure, first and second-level factor loadings, and intercepts are identical in both genders. It is true, however, that by constraining the item intercepts, the model fit worsened significantly. Even so, as the CFI loss was very close to .01, we considered accepting the invariance. Future studies could develop a detailed analysis of the differential items' performance to identify whether this loss of fit is due to the poor performance of any particular item.

Finally, concerning the limitations of this study, it is worth mentioning that it is focused exclusively on secondary school students. Thus, the generalizations of the results to younger-age or university students has not been demonstrated. We could anticipate that if reliability problems of the resilience dimension are related to difficulties in understanding a complex statement, they could increase in younger samples. Thus, an invariance study considering university stages would be interesting.

In conclusion, it should be noted that the instrument presents certain psychometric limitations that question its use in the sample to assess some of its specific dimensions: resilience and optimism. Although it could be useful for a general assessment of Psychological Capital, given the evidence of the existence of this second-order factor and the gender invariance of the scale, other alternatives should be explored if an assessment of each of its dimensions is desired.

References

Avey, J. B. (2014). The left side of psychological capital. *Journal of Leadership and Organizational Studies, 21*(2), 141-149. http://doi.org/10.1177/1548051813515516

Avey, J. B., Avolio, B. J., & Luthans, F. (2011). Experimentally analyzing the impact of leader positivity on follower positivity and performance. *The Leadership Quarterly, 22*(2), 282-294. http://doi.org/10.1016/j.leaqua.2011.02.004

Azanza, G., Domínguez, A. J., Moriano, J. A., & Molero, F. J. (2014). Capital psicológico positivo. Validación del cuestionario PCQ en España. *Anales de Psicología, 30*(1), 294-301. http://doi.org/10.6018/analesps.30.1.153631

Bandura, A. (1997). *Self-efficacy: The exercise of control*. W. H. Freeman.

Carmona-Halty, M., Salanova, M., Llorens, S., & Schaufeli, W. B. (2019). How psychological capital mediates between study-related positive emotions and academic performance. *Journal of Happiness Studies, 20*(2), 605-617. http://doi.org/10.1007/s10902-018-9963-5

Chen, F. F., Sousa, K. H., & West, S. G. (2005). Teacher´s corner: Testing measurement invariance of second order factor models. *Structural Equation Modeling: A Multidisciplinary Journal, 12*(3), 471-492. http://doi.org/10.1207/s15328007sem1203_7

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling: A Multidisciplinary Journal, 9*(2), 233-255. http://doi.org/10.1207/s15328007sem0902_5

Clapp-Smith, R., Vogelgesang, G. R., & Avey, J. B. (2009). Authentic leadership and positive psychological capital: The mediating role of trust at the group level of analysis. *Journal of Leadership and Organizational Studies, 15*(3), 227-240. http://doi.org/10.1177/1548051808326596

Datu, J. A. D., King, R. B., & Valdez, J. P. M. (2018). Psychological capital bolsters motivation, engagement, and achievement: Cross-sectional and longitudinal studies. *Journal of Positive Psychology, 13*(3), 260-270. http://doi.org/10.1080/17439760.2016.1257056

Dimitrov, D. M. (2010). Testing for factorial invariance in the context of construct validation. *Measurement and Evaluation in Counseling and Development, 43*(2), 121-149. http://doi.org/10.1177/0748175610373459

Djourova, N., Rodriguez, I., & Lorente-Prieto, L. (2019). Validation of a modified version of the Psychological Capital Questionnaire (PCQ12) in Spain. *Revista Interamericana de Psicología Ocupacional, 37*(2), 93-106. http://revista.cincel.com.co/index.php/RPO/article/view/228

Little, T. D. (1997). Mean and covariance structures (MACS) analyses of cross-cultural data: Practical and theoretical issues. *Multivariate Behavioral Research, 32*(1), 53-76. http://doi.org/10.1207/s15327906mbr3201_3

Luthans, F., Avolio, B. J., Avey, J. B., & Norman, S. M. (2007). Positive psychological capital: Measurement and relationship with performance and satisfaction. *Personnel Psychology, 60*(3), 541-572. http://doi.org/10.1111/j.1744-6570.2007.00083.x

Luthans, B. C., Luthans, K. W., & Jensen, S. M. (2012). The impact of business school students' psychological capital on academic performance. *Journal of Education for Business, 87*(5), 253-259. http://doi.org/10.1080/08832323.2011.609844

Luthans, F., & Youssef, C. M. (2004). Human, social, and now positive psychological capital management: Investing in people for competitive advantage. *Organizational Dynamics, 33*(2), 143-160. http://doi.org/10.1016/j.orgdyn.2004.01.003

Luthans, F., Youssef-Morgan, C. M., & Avolio, B. J. (2015). *Psychological capital and beyond*. Oxford University Press.

Marsh, H. W., Hau, K-. T., & Wen, Z. (2004). In search of golden rules: Comment on hypothesis-testing approaches to setting cutoff values for fit indexes and dangers in overgeneralizing Hu and Bentler's (1999) findings. *Structural Equation Modeling: A Multidisciplinary Journal, 11*(3), 320-341. http://doi.org/10.1207/s15328007sem1103_2

Martínez, I. M., Meneghel, I., Carmona-Halty, M., & Youssef-Morgan, C. M. (2021). Adaptation and validation to Spanish of the Psychological Capital Questionnaire-12 (PCQ-12) in academic contexts. *Current Psychology, 40*(7), 3409-3416. http://doi.org/10.1007/s12144-019-00276-z

Muthén, L. K., & Muthén, B. O. (2017). Mplus (8.9). [software de cómputo]. https://www.statmodel.com

Peña-Contreras, E. K., Lima-Castro, S. E., Arias-Medina, W. P., Bueno-Pacheco, G. A., Aguilar-Sizer, M. E., & Cabrera-Vélez, M. M. (2020). Propiedades psicométricas de la Escala Breve de Resiliencia (BRS) en el contexto ecuatoriano. *Revista Evaluar, 20*(3), 83-98. http://doi.org/10.35670/1667-4545.v20.n3.31715

Raykov, T., & Marcoulides, G. A. (2012). Evaluation of validity and reliability for hierarchical scales using latent variable modeling. *Structural Equation Modeling: A Multidisciplinary Journal, 19*(3), 495-508. http://doi.org/10.1080/10705511.2012.687675

Rego, A., Sousa, F., Marques, C., & Pina e Cunha, M. (2012). Authentic leadership promoting employees' psychological capital and creativity. *Journal of Business Research, 65*(3), 429-437. http://doi.org/10.1016/j.jbusres.2011.10.003

Schönfeld, F. S., & Mesurado, B. (2020). Adaptación del Cuestionario de Capital Psicológico al ámbito educativo en una muestra argentina. *Propósitos y Representaciones, 8*(1), e315. http://doi.org/10.20511/pyr2020.v8n1.315

Slåtten, T., Lien, G., Evenstad, S. B. N., & Onshus, T. (2021). Supportive study climate and academic performance among university students: The role of psychological capital, positive emotions and study engagement. *International Journal of Quality and Service Sciences, 13*(4), 585-600. http://doi.org/10.1108/IJQSS-03-2020-0045

Stajkovic, A. D., & Luthans, F. (2003). Behavioral management and task performance in organizations: Conceptual background, meta-analysis, and test of alternative models. *Personnel Psychology, 56*(1), 155-194. http://doi.org/10.1111/j.1744-6570.2003.tb00147.x

Tomás, J. M., Gutiérrez, M., Georgieva, S., & Hernández, M. (2020). The effects of self-efficacy, hope, and engagement on the academic achievement of secondary education in the Dominican Republic. *Psychology in the Schools, 57*(2), 191-203. http://doi.org/10.1002/pits.22321

Tomás, J. M., Martínez-Gregorio, S., & Oliver, A. (2022). Bayesian confirmatory factor analysis of the Psychological Capital PCQ-12 Scale. *European Journal of Psychological Assessment*. http://doi.org/10.1027/1015-5759/a000738

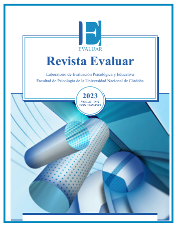
Vuyk, M. A., & Codas, G. (2019). Validación de la Escala de Esperanza Disposicional para Adultos en Paraguay. *Revista Evaluar, 19*(1), 59-71. http://doi.org/10.35670/1667-4545.v19.n1.23880

Wang, J., & Wang, X. (2012). *Structural equation modeling: Applications using Mplus*. John Wiley & Sons. http://doi.org/10.1002/9781118356258

EVALUAR

2023, Vol. 23, No. 2
ISSN 1667-4545
Recuperado de https://revistas.unc.edu.ar/index.php/revaluar

Laboratorio de Evaluación Psicológica y Educativa
Facultad de Psicología - Universidad Nacional de Córdoba



Assessment of Breastfeeding Engagement in Postpartum Women with the UWES-17 scale

Evaluación del compromiso en el amamantamiento de mujeres puérperas con escala UWES-17

Introduction
Methods
Results
Discussion
References

Mónica L. Brizuela *[1], Agustín R. Miranda[2], Lucía Sogno[1], Luisina Malpassi[1], Elio A. Soria[3], Silvana V. Serra[1]

1 - Escuela de Fonoaudiología, Facultad de Ciencias Médicas, Universidad Nacional de Córdoba, Argentina.
2 - Montpellier Interdisciplinary Center on Sustainable Agri-food Systems (MoISA), University of Montpellier, CIRAD, CIHEAM-IAMM, INRAE, Institut Agro, IRD, Montpellier, France. 911 Avenue d'Agropolis, Cedex 5, 34394, Montpellier, France.
3 - Consejo Nacional de Investigaciones Científicas y Técnicas, CONICET, INICSA, Blvd. de la Reforma s/n, Ciudad Universitaria, 5016, Córdoba, Córdoba, Argentina.

Recibido: 28/02/2023 **Revisado:** 20/04/2023 **Aceptado:** 19/05/2023

Abstract

Promotion and monitoring of breastfeeding require reliable and valid instruments that allow studying the engagement of itself. Thus, the aim of this study was to analyze the psychometric properties of the Utrecht Work Engagement Scale (UWES-17) for the assessment of engagement in breastfeeding practices among 324 postpartum Argentinian women. The UWES-17, Breastfeeding Self-Efficacy Scale, Postpartum Depression Scale, and a sociodemographic questionnaire were applied. Moreover, reliability, validity, dimensionality, sensitivity, and specificity were analyzed. The UWES-17 demonstrated adequate levels of internal consistency, and its three-dimensional structure was confirmed. Bifactorial analysis supported its usage, and the model verified its external validity. The results validate the UWES-17 as a valid and reliable tool for assessing breastfeeding engagement, thus making it suitable for implementation in clinical and scientific contexts to support interdisciplinary approaches to breastfeeding.

Keywords: *breastfeeding, puerperal women, psychometrics, maternal and child health, self-report*

Resumen

La promoción y el seguimiento de la lactancia materna necesita contar con instrumentos fiables y válidos que permitan estudiar el grado de compromiso con el amamantamiento. El objetivo fue analizar las propiedades psicométricas de la Escala de Compromiso de Utrecht (UWES-17) para evaluar el compromiso en la práctica del amamantamiento en 324 mujeres puérperas argentinas. Se utilizaron los instrumentos UWES-17, Escala de Autoeficacia para la Lactancia Materna, Escala de Depresión Postparto y cuestionario sociodemográfico. Se analizaron la fiabilidad, validez, dimensionalidad, sensibilidad y especificidad. El UWES-17 mostró niveles adecuados de consistencia interna y se confirmó su estructura tridimensional. El análisis bifactorial confirmó su utilidad y el modelo comprueba su validez externa. Los hallazgos confirman que el UWES-17 es un instrumento válido y fiable para la medición del compromiso en el amamantamiento, ya que puede ser utilizado en el ámbito clínico y científico para el abordaje interprofesional de la lactancia humana.

Palabras clave: *lactancia materna, mujeres puérperas, psicometría, salud materno infantil, autoinforme*

***Contact:** monica.brizuela@unc.edu.ar. Blvd. de la Reforma, Ciudad Universitaria. C.P: 5014, Córdoba. Argentina. Teléfono: + 54 9 3515517009.

How to cite: Brizuela, L. M., Miranda, R. A., Sogno L., Malpassi, L., Soria A. E., & Serra, V. S. (2023): Assessment of Breastfeeding Engagement in Postpartum Women with the UWES-17 scale (ABPW). *Revista Evaluar*, *23*(2), 28-39. Retrieved from https://revistas.unc.edu.ar/index.php/revaluar

Participaron en la edición de este artículo: Vanesa Mariela Toledo, Andrea Suárez, Eugenia Barrionuevo, Abigail Peréz, Gloria Rebeca Nieve, Florencia Ruiz, Jorge Bruera.

Introduction

Postpartum is a transitional period accompanied by significant changes and challenges that can affect maternal and infant well-being (Carrizo et al., 2020). In recent years, the study of the development of positive attitudes, skills, and experiences in maternal care has been addressed by the health sciences (Corno et al., 2019). Studies on positive psychology in the postpartum period aim to identify factors which promote optimal functioning and the development of personal resources. By increasing and promoting positive resources, it is possible to successfully counteract negative experiences and psychological disorders during this vulnerable stage (Corno et al., 2019).

Engagement is one of the most well-known theoretical constructs in the field of psychology and it is defined as a positive, persistent, emotional, and cognitive state related to adherence and sustainability in a task (Schaufeli et al., 2002). Maternal engagement refers to the mental state of a woman during infant care tasks (Provenzi et al., 2017). The authors of the theory concluded that engagement consists of three closely related components: vigor, dedication and absorption (Schaufeli et al., 2002). Vigor refers to high levels of energy and resilience while performing a task, that is, willingness to make an effort and persist even when facing difficulties. Dedication involves a high degree of involvement in the task, a sense of its significance, as well as experiencing enthusiasm, pride, inspiration, and challenge. Absorption describes full concentration and task enjoyment. From this perspective, it is known that appropriate maternal engagement is related to positive outcomes, including facilitating attachment and psycho-emotional development, promoting maternal mental well-being, and supporting the practice of breastfeeding, among other aspects (Carrizo et al., 2020; Corno et al., 2019).

Breastfeeding, as a cultural practice inherent to human beings that requires the postpartum person's commitment, can be studied using self-report instruments (Girard et al., 2016). Postpartum women's engagement to breastfeeding is related to involvement, behavior, personal initiative, performance, and quality in the activity of breastfeeding (Wouk et al., 2020). Relationships between engagement and better mental and physical health can increase motivation, self-efficacy, optimism, and self-esteem, making it opportune to measure this construct in relation to breastfeeding (Wouk et al., 2020). Furthermore, the World Health Organization (2001) declared that both clinical and population-based research are a priority to achieve long-term global goals to increase the engagement of breastfeeding. Therefore, it is necessary for healthcare professionals and researchers to have reliable and valid instruments to assess lactating women's level of engagement in order to develop timely strategies to increase the percentage of breastfeeding.

Consequently, the present study aimed to examine the psychometric properties of the UWES-17 instrument for the assessment of maternal engagement in breastfeeding practices among Argentinian lactating women. Given the need for instruments that provide self-perceived information, there is a fundamental need to adapt and validate them in the local context in which they are intended to be used (Olivera et al., 2023).

Methods

Participants

An analytical cross-sectional study was carried out, which involved the administration of online self-report questionnaires to 324 postpartum women in Argentina. The inclusion criteria were: adults (≥ 18 years) residing in Argentina who were breastfeeders in the postpartum stage (first twelve months). Participants signed an informed consent before voluntarily participating. This research was approved by the corresponding Research Ethics Committee (REPIS-3177).

Instruments

Breastfeeding engagement. The Spanish version of the Utrecht Work Engagement Scale (UWES-17) was used, consisting of 17 items that assess the subscales of vigor, dedication, and absorption (Wouk et al., 2020). Participants rated the frequency at which they have felt the described statements in each item using a Likert-type scale with seven options ranging from never - *not at all* (0) to *always* - *every day* (6). Schaufeli and Bakker (2004) reported Cronbach's alpha values for the UWES scale ranging from .80 to .90.

For the current study, experts in breastfeeding collaborated in adapting the items to the breastfeeding practice. The Spanish version of the 17-item UWES was modified following the procedure by Guillén and Martínez-Alvarado (2014) for adapting the UWES in non-work contexts. Once the adaptation of the items was completed, a pilot test was conducted with 65 women, and the necessary adjustments were made to obtain the final version of the questionnaire.

Breastfeeding self-efficacy. Women's confidence in breastfeeding was assessed using the Spanish version of the Breastfeeding Self-Efficacy Scale-Short form (BSES-SF). The BSES-SF is valid for identifying women who experienced difficulties in breastfeeding and has been used in evaluating support interventions (Oliver-Roig et al., 2012). This instrument consists of 14 positively framed items with the phrase *I can always* rate on a 5-point Likert scale. The total score ranges from 14 to 70, with higher scores that indicate higher levels of self-efficacy in breastfeeding. Scores above 50 are indicative of adequate self-efficacy in breastfeeding (sensitivity > 70% and specificity > 50%) (Nanishi et al., 2015). In this study, the reliability was good (alpha = .87).

Mood. The 7-item Spanish version of the Postpartum Depression Screening Scale (PDSS) was used to assess suggestive signs of depressive disorders during the postpartum period (Le et al., 2010). Each item was rated on a 5-point Likert scale. The scores were transformed into a scale ranging from 7 to 35, with higher scores that indicate higher levels of depression (Miranda et al., 2021). In this study, the alpha showed an acceptable value of .83.

Sociodemographic and health variables. An ad hoc questionnaire was designed to collect sociodemographic data (age, relationship status, years of education, employment, access to healthcare) and gynaecological-obstetric data (type of delivery, parity, number of pregnancies, postpartum period, gestational characteristics and type of breastfeeding).

Statistical analysis

Statistical analyses were performed using Stata 17 software and included confirmatory factor analysis (CFA) with structural equation modeling techniques (SEM) on the 17 items with the *a priori* identified dimensions (three-dimensional model), compared with the unifactorial model. Traditional goodness-of-fit indices were calculated, and reliability and validity (convergent and divergent) were assessed. Subsequently, a bifactorial model was tested to determine if the measure was sufficiently unidimensional to support the use of a total score, while still considering the multidimensionality found (Miranda et al., 2020). Internal consistency was evaluated using Cronbach's alpha (α), with acceptable values ranging from .60 to .95. Additionally, item correlations (*r* coefficients) were calculated to detect potential redundancy among them ($r > .80$) (Miranda et al., 2020). Pearson correlation matrices (*r*) were estimated between the UWES-17 and the BSES-SF to determine convergent validity, while the PDSS was used to study divergent validity. Furthermore, the three questionnaires were integrated into a SEM model to assess nomological validity. Lastly, the sensitivity and specificity of the UWES-17 were studied by plotting ROC curves, using an adequate self-efficacy (BSES-SF > 50 points) as the parameter. The area under the curve (AUC) was estimated, considering that a higher AUC indicates better discrimination in identifying women with self-efficacy: non-discriminatory (< .60), fair (.60 to .69), acceptable (.70 to .79), excellent (.80 to .90), and outstanding (> .90).

Results

Most women under 35 years old (82%) were in a relationship (96%), and had received at least 12 years of formal education (92.8%). Furthermore, 78% were employed and 86% had private healthcare coverage. Regarding obstetrical and gynaecological data, 56% of the participants were primiparous, 51% were multigravida, and 27% had experienced a previous pregnancy loss. Around 73% were recruited during the first 6 months of postpartum, and exclusive breastfeeding was practiced by 64% of the sample.

Goodness-of-fit measures are shown in Table 1 for the CFA of the unidimensional and tridimensional models of the UWES-17 for lactation. The indices confirmed that the questionnaire has a tridimensional structure. All indices reached the recommended values: $\chi^2/df = 2.74$, CFI = .91, TLI = .90, RMSEA = 0.07, SRMR = 0.05, CD = 0.92, and decreasing AIC and BIC. Figure 1 displays the structural equation models. Furthermore, the CFA supported a bifactorial structure, which demonstrated a superior fit ($\chi^2/df = 2.62$, CFI = .92, TLI = .90, RMSEA = 0.06, SRMR = 0.05, CD = .99).

Table 1
Fit indices for the confirmatory factor analysis models.

	Expected values	Unidimensional model	Three dimensional model	Bifactor analysis
χ^2/df	≤ 3.00	3.61	2.74	2.62
RMSEA	< 0.08	0.09	0.07	0.06
CFI	≥ .90	.86	.91	.92
TLI	≥ .90	.84	.90	.90
SRMR	≤ 0.08	0.06	0.05	0.05
CD	≥ .95	.93	.92	.99
AIC	The lower the better	18433.12	18331.26	18311.66
BIC		18622.69	18554.28	18571.85

Note. χ^2/df = chi-square to degree of freedom ratio; RMSEA = root mean square error of approximation; CFI = comparative fit index; TLI = Tucker–Lewis index; SRMR = standardized root mean square residual; CD = coefficient of determination; AIC = Akaike information criterion; BIC = Bayesian information criterion.

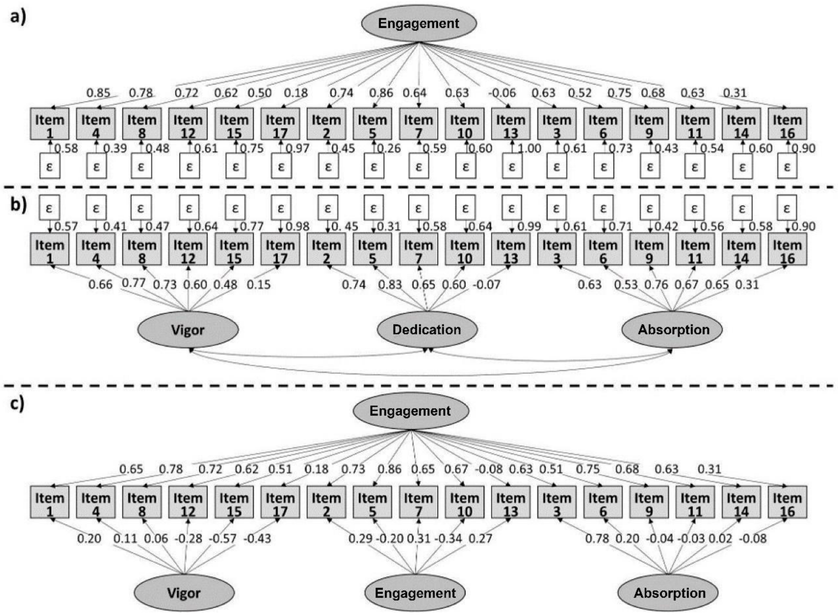


Figure 1
Structural equations of the unidimensional model (a), tridimensional model (b), and bifactorial model (c) of the UWES-17 for breastfeeding engagement.

Regarding dimensionality, the correlations were below .90. The questionnaire achieved very good reliability for the entire instrument, with α = .89. The factor with the highest α value was vigor (α = .76), followed by absorption (α = .75), and dedication (α = .63). Additionally, the α coefficients did not significantly change after the removal of each item (Table 2).

Table 2
Analysis of the reliability of the UWES-17 for breastfeeding engagement.

Item	IT	IR	Alpha
1	.64	.58	.88
2	.73	.68	.87
3	.66	.60	.80
4	.76	.71	.87
5	.81	.78	.87
6	.57	.49	.88
7	.66	.60	.88
8	.73	.67	.87
9	.74	.70	.87
10	.63	.59	.88
11	.68	.65	.88
12	.67	.62	.87
13	.15	.02	.90
14	.68	.62	.87
15	.57	.50	.88
16	.39	.30	.89
17	.27	.20	.89

Note. IT = item-test correlation; IR = item-rest correlation; α = alpha after item deletion.

Subsequently, the correlation between the UWES-17 and PDSS-SF was analyzed to assess divergent validity. The PDSS was inversely correlated with the total score of the UWES-17 (*r* = -.19, *p* < .01), vigor (*r* = -.25, *p* < .01), dedication (*r* = -.13, *p* < .01), and absorption (*r* = -.14, *p* < .01). As for convergent validity, the UWES-17 showed significant positive correlations with the BSES-SF (*r* = .55, *p* < .01). It also correlated with vigor (*r* = .61, *p* < .01), dedication (*r* = .33, *p* < .01), and absorption (*r* = .51, *p* < .01). Furthermore, a theoretical model was designed to study the multivariate relationship of the questionnaires through SEM analysis (Figure 2).

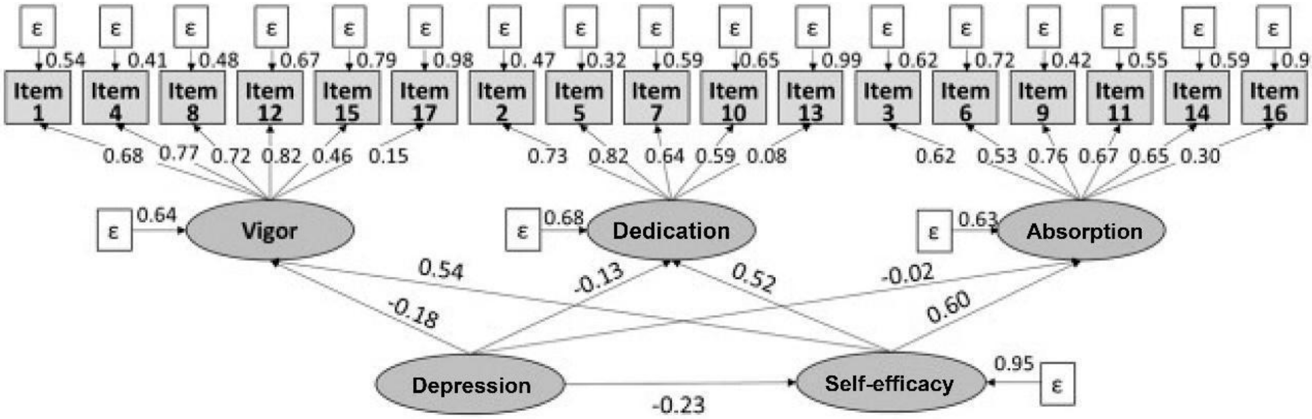


Figure 2
Nomological validity of the UWES-17 for breastfeeding engagement.

The results indicated a good fit between data and structural model: $\chi^2/df = 2.73$, RMSEA = 0.07, CFI = .91, TLI = .88, SRMR = 0.06, CD = .29. Therefore, the SEM provided evidence of nomological validity for the instrument. The direct effects of postpartum depression (PDSS) and self-efficacy on the components of engagement (UWES-17) are shown in Figure 2. The results showed that postpartum depression negatively predicted vigor (β = -.18, *p* < .01) and dedication (β = -.13, *p* = .02) in breastfeeding practice. On top of that, it was found that self-efficacy in breastfeeding had positive effects on vigor (β = .54, *p* < .01), dedication (β = .52, *p* < .01) and absorption (β = .60, *p* < .01).

The standardized scores of the UWES-17 stratified by parity are presented as mean, standard deviation, and percentile ranges in Table 3. The scores were similar for both groups, and the means of the participants in the UWES-17 subscales were above the 25th percentile.

Finally, a cutoff point of 67 demonstrated the optimal balance between sensitivity (70.94%) and specificity (69.01%) (Figure 3a). A ROC curve was calculated (Figure 3b) to identify the relative sensitivity/specificity of the UWES-17 instrument compared to the comparison instrument (BSES-SF). The AUC for the UWES-17 was 0.74 (SE = 0.03; *p* < .01).

Table 3
Means, standard deviations, and percentiles of the total UWES and its subscales.

Parity	Score	n	M	SD	P5	P10	P25	P50	P75	P90	P95
Multiparous	Vigor	128	26.06	7.18	13	16	21	27	32	34	35
	Dedication	128	21.01	6.09	10	13	18	22	26	29	29
	Absorption	128	23.91	6.74	11	15	20	25	29	32	33
	UWES total	128	70.98	18.39	36	44	59	76	86	90	94
Primiparous	Vigor	177	25.82	7.21	14	16	21	27	31	35	36
	Dedication	177	20.66	5.34	11	13	17	22	24	28	29
	Absorption	177	23.38	7.16	10	13	19	25	29	32	33
	UWES total	177	69.86	17.76	39	45	58	71	84	90	94

Note. M = mean; SD = standard deviation; P = percentile.

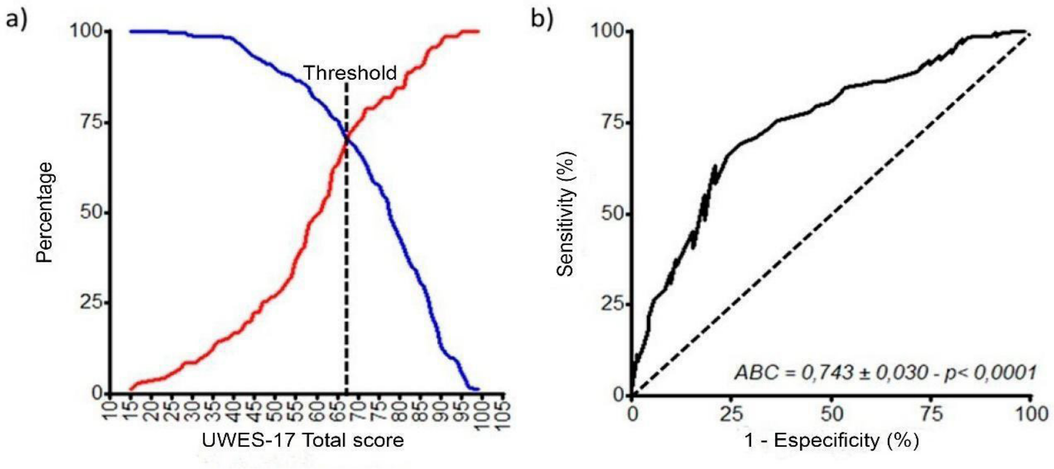


Figure 3
Sensitivity-specificity curves (a) and receiver operating characteristic (ROC) curves (b) of the UWES-17 questionnaire.

Discussion

This study aimed to examine the psychometric properties of the UWES-17 scale for assessing breastfeeding engagement in Argentinian women. The UWES-17 scale demonstrated satisfactory levels of internal consistency, structural validity, convergent validity, divergent validity, and nomological validity, making it useful for assessing the degree of breastfeeding engagement. The evaluation of validity and reliability is one of the most fundamental aspects in the development, evaluation, and use of instruments (Chan, 2014). Furthermore, having valid and reliable instruments is crucial for healthcare professionals to identify women with lower engagement to breastfeeding and implement interventions that ensure the promotion and maintenance of lactation (Chambers et al., 2007).

Regarding reliability, the UWES-17 showed adequate levels of internal consistency, indicating that the items are coherent with each other and measure the same construct (Adamson & Prion, 2013). The alpha coefficients were satisfactory both overall and for each of the factors. There was no substantial improvement in internal consistency after removing items, and the analysis of item-item correlations did not identify ambiguities, indicating that all 17 items should be retained. These results are consistent with previous studies, which demonstrate that the UWES-17 is a reliable instrument for measuring engagement in work (Schaufeli et al., 2002), academic (Wickramsinghe et al., 2018) and sports activities (Guillén & Martínez-Alvarado, 2014).

Traditionally, this instrument has shown a three-dimensional structure: vigor, dedication, and absorption (Schaufeli et al., 2002; Gómez-Garbero et al., 2019). In this study, its three-dimensionality was confirmed in the adapted version for breastfeeding, which yielded satisfactory goodness-of-fit statistics. In accordance with current recommendations, a bifactor analysis was carried out to ascertain whether the UWES-17 could identify the presence of engagement in breastfeeding as a latent dimension underlying the total score of the UWES-17 (Morin et al., 2020). Bifactor analysis is used to determine if a measure is sufficiently unidimensional to support the use of a total score while still accounting for multidimensionality (Reise et al., 2007). Therefore, UWES-17 scores can be interpreted both from the scores of its factors and the overall score, which is relevant for a comprehensive approach to breastfeeding and the development of specific guidelines that consider various postpartum health scenarios (Smorti et al., 2020). Following the logical reasoning proposed by the authors of the engagement theory (Schaufeli et al., 2002), committed postpartum women demonstrate vigor, dedication, and absorption in breastfeeding, where vigor is manifested by energy, mental resilience and willingness to make efforts to complete breastfeeding. Faced with challenges, these women remain persistent. A committed postpartum woman also perceives her breastfeeding practice as meaningful, addressing tasks with care and dedication. Additionally, engagement is characterized by absorption; this means that postpartum women are fully concentrated on the task of breastfeeding, where time seems to pass quickly, and they may even experience some difficulty in disengaging from it (Schaufeli et al., 2002).

Once the reliability of the UWES-17 and its dimensionality were confirmed, the analysis of convergent and divergent validity was carried out by contrasting it with instruments that directly (convergent validity) and indirectly (divergent validity) assess related theoretical dimensions. When evaluating divergent validity, the UWES-17 showed an inverse correlation with postpartum depression. These findings are consistent with previous research on postpartum depression and breastfeeding, which indicate that postpartum depression symptoms can compromise breastfeeding (Avilla et al., 2020). Convergent validity was confirmed by calculating the associations between the UWES-17 and self-efficacy in breastfeeding, which were directly correlated. This result is expected since self-efficacy refers to the belief in one's own capabilities to organize and execute the necessary actions to achieve certain goals, in this case, breastfeeding (Ghasemi et al., 2019). Self-efficacy is a cognitive resource necessary for establishing adequate engagement, as a strong sense of self-efficacy can contribute to achieving a balance between the various demands faced by women during lactation (Miranda et al., 2020). Additionally, self-efficacy indicates a person's motivation and willingness to make consistent efforts in line with their abilities (Ghasemi et al., 2019). Self-efficacy is a key motivational belief that has been conceptually and empirically linked to self-regulatory beliefs. Moreover, self-regulatory efficacy refers to beliefs about using self-regulated learning processes, such as goal setting, self-monitoring, strategy use, self-evaluation, and self-reaction (Zimmerman et al., 2005).

Furthermore, all three questionnaires were included in a SEM model to test nomological validity, which confirmed the associations between them with an adequate level of fit. Nomological validity functions as an additional approach to evaluate the construct validity of a questionnaire, ensuring that the observed correlations between variables are aligned with theoretical or hypothetical relationships (Hair Jr et al., 2020). In the present study, the verification of nomological validity indicates that the structure of the UWES-17 aligns with the theoretical assertions in the literature, where the level of engagement depends on a person's psychological capital (Schaufeli, 2013). These results highlight the need to take care of women's mental health during the postpartum period. This is the reason it is important to provide specific support for breastfeeding to women with psychological difficulties, as previous research has emphasized (Borra et al., 2015). In this regard, lactation consultations have been shown to improve breastfeeding self-efficacy and maternal mental health (Chrzan-Dętkoś et al., 2021).

Finally, standardized scores of the UWES-17 stratified by parity are presented since it is one of the main factors predicting the initiation and maintenance of breastfeeding (Chang et al., 2019). Similar to previous findings where a score below 70 identified employees at psychological risk (Roelen et al., 2015), the cutoff point of 67 demonstrates the optimal trade-off between sensitivity (70.94%) and specificity (69.01%). Therefore, it is recommended to use a UWES-17 score of 67 as a cutoff point when screening to predict the level of breastfeeding engagement. Scores equal to or below 67 can be taken as indicators of the need for interventions to support breastfeeding and identify possible risk factors, such as postpartum depression.

Lastly, it is necessary to acknowledge some limitations. We recommend that future research studies breastfeeding engagement longitudinally in order to identify risk factors and promoters. Additionally, conducting research in clinical populations, such as women and/or infants with health conditions that hinder breastfeeding, is suggested to improve applicability. Despite these limitations, our findings contribute to the existing evidence and raise new questions.

In conclusion, the UWES-17 is a reliable and valid questionnaire for assessing breastfeeding engagement. Our findings suggest that the UWES-17 is an appropriate tool for identifying women at risk of suboptimal breastfeeding outcomes and can provide a strategy to recognize those who may benefit from breastfeeding interventions. Having appropriate tools to comprehensively address women's health during the postpartum period is crucial for healthcare professionals and scientists, as establishing breastfeeding engagement serves as a long-term indicator of lactation sustainability.

References

Adamson, K. A., & Prion, S. (2013). Reliability: Measuring internal consistency using Cronbach's α. *Clinical Simulation in Nursing, 9*(5), e179-e180. https://doi.org/10.1016/j.ecns.2012.12.001

Avilla, J. C., Giugliani, C., Bizon, A. M. B. L., Martins, A. C. M., Senna, A. F. K., & Giugliani, E. R. J. (2020). Association between maternal satisfaction with breastfeeding and postpartum depression symptoms.

Plos One, 15(11), e0242333. https://doi.org/10.1371/journal.pone.0242333

Borra, C., Iacovou, M., & Sevilla, A. (2015). New evidence on breastfeeding and postpartum depression: The importance of understanding women's intentions. *Maternal and Child Health Journal, 19*(4), 897-907. https://doi.org/10.1007/s10995-014-1591-z

Carrizo, E., Domini, J., Quezada, R. Y. J., Serra, S. V., Soria, E. A., & Miranda, A. R. (2020). Variaciones del estado cognitivo en el puerperio y sus determinantes: Una revisión narrativa. *Ciência & Saúde Coletiva, 25*(8), 3321-3334. https://doi.org/10.1590/1413-81232020258.26232018

Chambers, J. A., McInnes, R. J., Hoddinott, P., & Alder, E. M. (2007). A systematic review of measures assessing mothers' knowledge, attitudes, confidence and satisfaction towards breastfeeding. *Breastfeeding Review, 15*(3), 17-25.

Chan, E. K. H. (2014). Standards and guidelines for validation practices: Development and evaluation of measurement instruments. In B. D. Zumbo & E. K. H. Chan (Eds.), *Validity and validation in social, behavioral, and health sciences* (9-24). Springer International Publishing.

Chang, P. C., Li, S. F., Yang, H. Y., Wang, L. C., Weng, C. Y., Chen, K. F., & Fan, S. Y. (2019). Factors associated with cessation of exclusive breastfeeding at 1 and 2 months postpartum in Taiwan. *International Breastfeeding Journal, 14*(1), 1-7. https://doi.org/10.1186/s13006-019-0213-1

Chrzan-Dętkoś, M., Walczak-Kozłowska, T., Pietkiewicz, A., & Żołnowska, J. (2021). Improvement of the breastfeeding self-efficacy and postpartum mental health after lactation consultations - Observational study. *Midwifery, 94*, 102905. https://doi.org/10.1016/j.midw.2020.102905

Corno, G., Espinoza, M., & Baños, R. M. (2019). A narrative review of positive psychology interventions for women during the perinatal period. *Journal of Obstetrics and Gynaecology, 39*(7), 889-895. https://doi.org/10.1080/01443615.2019.1581735

Ghasemi, V., Simbar, M., Banaei, M., Naz, M. S. G., Jahani, Z., & Nazem, H. (2019). The effect of interventions on breastfeeding self-efficacy by using Bandura's theory in Iranian mothers: A systematic review. *International Journal of Pediatrics, 7*(8), 9939-9954. https://ijp.mums.ac.ir

Girard, L. C., Côté, S. M., de Lauzon-Guillain, B., Dubois, L., Falissard, B., Forhan, A., Doyle, O., Bernard, J. Y., Heude, B., Saurel-Cubizolles, M. J., Kaminski, M., Boivin, M., Tremblay, R. E., & Eden Mother-Child Cohort Study Group. (2016). Factors associated with breastfeeding initiation: A comparison between France and French-speaking Canada. *PloS One, 11*(11), e0166946. https://doi.org/10.1371/journal.pone.0166946

Gómez-Garbero, L., Labarthe, J., Ferreira-Umpiérrez, A., & Chiminelli-Tomás, V. (2019). Evaluación del engagement en trabajadores de la salud en Uruguay a través de la Escala Utrecht de Engagement en el Trabajo (UWES). *Ciencias Psicológicas, 13*(2), 305-316. https://doi.org/10.22235/cp.v13i2.1888

Guillén, F., & Martínez-Alvarado, J. R. (2014). The sport engagement scale: An adaptation of the Utrecht Work Engagement Scale (UWES) for the sports environment. *Universitas Psychologica, 13*(3), 975-984. https://doi.org/10.11144/Javeriana.UPSY13-3.sesa

Hair Jr, J. F., Howard, M. C., & Nitzl, C. (2020). Assessing measurement model quality in PLS-SEM using confirmatory composite analysis. *Journal of Business Research, 109*, 101-110. https://doi.org/10.1016/j.jbusres.2019.11.069

Le, H. N., Perry, D. F., & Ortiz, G. (2010). The Postpartum Depression Screening Scale-Spanish version: Examining the psychometric properties and prevalence of risk for postpartum depression. *Journal of Immigrant and Minority Health, 12*(2), 249-258. https://doi.org/10.1007/s10903-009-9260-9

Miranda, A. R., Scotta, A. V., Cortez, M. V., & Soria, E. A. (2021). Triggering of postpartum depression and insomnia with cognitive impairment in Argentinian women during the pandemic COVID-19 social isolation in relation to reproductive and health factors. *Midwifery, 102*, 103072. https://doi.org/10.1016/j.midw.2021.103072

Miranda, A. R., Scotta, A. V., Méndez, A. L., Serra, S. V., & Soria, E. A. (2020). Public sector workers' mental health in Argentina: Comparative psychometrics of the Perceived Stress Scale. *Journal of Preventive Medicine and Public Health, 53*(6), 429-438. https://doi.org/10.3961/jpmph.20.229

Morin, A. J. S., Myers, N. D., & Lee, S. (2020). Modern factor analytic techniques: Bifactor models, exploratory structural equation modeling (ESEM), and bifactor-ESEM. En G. Tenenbaum, R. C. Eklund (Eds), *Handbook of Sport Psychology* (1044-1073). https://doi.org/10.1002/9781119568124.ch51

Nanishi, K., Green, J., Taguri, M., & Jimba, M. (2015). Determining a cut-off point for scores of the Breastfeeding Self-Efficacy Scale-Short Form: Secondary data analysis of an intervention study in Japan. *PloS One, 10*(6), e0129698. https://doi.org/10.1371/journal.pone.0129698

Olivera, M., Prozzillo, P., & Simkin, H. A. (2023). Listado de Evaluación del Soporte Interpersonal (LESI-12): Evidencias de validez y confiabilidad en el contexto argentino. *Revista Evaluar, 23*(1), 1-11. https://revistas.unc.edu.ar/index.php/revaluar/issue

Oliver-Roig, A., d'Anglade-González, M. L., García-García, B., Silva-Tubio, J. R., Richart-Martínez, M., & Dennis, C. L. (2012). The Spanish version of The Breastfeeding Self-efficacy Scale-short form: Reliability and validity assessment. *International Journal of Nursing Studies, 49*(2), 169-173. https://doi.org/10.1016/j.ijnurstu.2011.08.005

Provenzi, L., Fumagalli, M., Bernasconi, F., Sirgiovanni, I., Morandi, F., Borgatti, R., & Montirosso, R. (2017). Very preterm and full-term infants' response to socio-emotional stress: The role of postnatal maternal bonding. *Infancy, 22*(5), 695-712. https://doi.org/10.1111/infa.12175

Reise, S. P., Morizot, J., & Hays, R. D. (2007). The role of the bifactor model in resolving dimensionality issues in health outcomes measures. *Quality of Life Research, 16*, 19-31. https://doi.org/10.1007/s11136-007-9183-7

Roelen, C. A. M., van Hoffen, M. F. A., Groothoff, J. W., de Bruin, J., Schaufeli, W. B., & van Rhenen, W. (2015). Can the Maslach Burnout Inventory and Utrecht Work Engagement Scale be used to screen for risk of long-term sickness absence? *International Archives of Occupational and Environmental Health, 88*(4), 467-475. https://doi.org/10.1007/s00420-014-0981-2

Schaufeli, W. B. (2013). What is engagement? In C. Truss, K. Alfes, R. Delbridge, A. Shantz, & E. Soane (Eds.), *Employee Engagement in Theory and Practice* (pp. 15-35). Routledge.

Schaufeli, W. B., & Bakker, A. B. (2004). Job demands, job resources, and their relationship with burnout and engagement: A multi-sample study. *Journal of Organizational Behavior, 25*(3), 293-315. https://doi.org/10.1002/job.248

Schaufeli, W. B., Salanova, M., González-Romá, V., & Bakker, A. B. (2002). The measurement of engagement and burnout: A two sample confirmatory factor analytic approach. *Journal of Happiness Studies: An Interdisciplinary Forum on Subjective Well-Being, 3*(1), 71-92. https://link.springer.com/journal/10902

Smorti, M., Testa, I., Gallese, M., Dotti, A., Ionio, C., Andreol, A., Zilioli, A., Pravettoni, G., Greco, A., Fenaroli, V., Nastasi, G., Giuntini, N., & Bonassi, L. (2020). Protect, promote and support: A warm chain of breastfeeding for oncological women-results from a survey of young Italian cancer mothers. *Ecancermedicalscience, 14*. https://doi.org/10.3332/ecancer.2020.1151

Wickramasinghe, N. D., Dissanayake, D. S., & Abeywardena, G. S. (2018). Validity and reliability of the Utrecht Work Engagement Scale-student version in Sri Lanka. *BMC Research Notes, 11*(1), 1-6. https://doi.org/10.1186/s13104-018-3388-4

World Health Organization. (2001). *Global strategy for in-*

fant and young child feeding. The optimal duration of exclusive breastfeeding. 54 World Health Assembly. Geneva. https://apps.who.int/iris/handle/10665/78801

Wouk, K., Tucker, C., Pence, B. W., Meltzer-Brody, S., Zvara, B., Grewen, K., & Stuebe, A. M. (2020). Positive emotions during infant feeding and breastfeeding outcomes. *Journal of Human Lactation, 36*(1), 157-167. https://doi.org/10.1177/0890334419845646

Zimmerman, B. J., Kitsantas, A., & Campillo, M. (2005). Evaluación de la autoeficacia regulatoria: Una perspectiva social cognitiva. *Revista Evaluar, 5*(1), 01-21. https://doi.org/10.35670/1667-4545.v5.n1.537

EVALUAR

2023, Vol. 23, No. 2
ISSN 1667-4545
Recuperado de https://revistas.unc.edu.ar/index.php/revaluar

Laboratorio de Evaluación Psicológica y Educativa
Facultad de Psicología - Universidad Nacional de Córdoba

Assessing Critical Thinking Skills: A Diagnosis of Elementary Students

Evaluación de las destrezas del pensamiento crítico: Un diagnóstico de los estudiantes de primaria

María Antonia Manassero-Mas [1] , Ángel Vázquez-Alonso * [2]

1 - Facultad de Psicología, Universidad de las Islas Baleares (España).
2 - Instituto de Investigación e Innovación Educativa, Universidad de las Islas Baleares (España).

Introduction
Method
Results
Discussion
References

Recibido: 28/02/2023   **Revisado:** 05/03/2023 **Aceptado:** 16/03/2023

Abstract

Critical thinking (CT) is a central aim of the 21st century education, although its research lacks consensus, has been unequal and lacking in primary evaluation issues and primary education. These weaknesses justify the aim of this study: evaluating primary students' mastery of CT skills. The quantitative methodology diagnoses six CT skills (prediction, logical reasoning, comparison, classification, decision-making, and problem-solving) through a 48-item test, which 655 sixth-graders completed. The students display an average global mastery of the CT items and skills, and the test's and skill scores' standardization are also presented. The comparison of boys and girls shows that girls perform better than boys on most test items. The diagnoses suggest an intermediate mastery of CT skills in students, where girls widely outperform boys, and also propose some educational implications.

Keywords: *assessment, critical thinking, skills, normative data in primary education, gender differences*

Resumen

El pensamiento crítico (PC) es un objetivo central de la educación del siglo XXI, aunque su investigación carece de consenso y ha sido desigual, y limitada en temas de evaluación en la educación primaria. Estas debilidades justifican el objetivo de este estudio: evaluar el dominio de las destrezas de PC en estudiantes de primaria. La metodología cuantitativa evalúa seis destrezas de CT (predicción, razonamiento lógico, comparación, clasificación, toma de decisiones y resolución de problemas) a través de un test de 48 ítems (alfa = .85), en el que participaron 655 estudiantes de sexto grado (11 años). Los estudiantes exhiben un dominio general promedio de los ítems y destrezas del pensamiento crítico. Además, se presentan las estandarizaciones de los resultados de los test y el puntaje obtenido en los ítems. La comparación de PC entre niños y niñas indica que las niñas obtienen mejores puntuaciones que los niños en la mayoría de los ítems. Con el estudio se concluye que existe un dominio intermedio del PC, que es ampliamente mejor en niñas, y se discuten algunas implicaciones educativas.

Palabras clave: *evaluación, pensamiento crítico, destrezas, baremación en primaria, diferencias de género*

***Correspond to:** Ángel Vázquez-Alonso. Dirección: Universidad de las Islas Baleares, Edificio Guillem Cifre de Colonya, Carretera de Valldemossa, km. 7.5, 07122, Palma, España. Teléfono: +34 971 17 30 75; FAX: 34 971 17 31 90. E-mail: angel.vazquez@uib.es

Author´s Note: Proyecto EDU2015-64642-R (AEI/FEDER, UE) financiado por la Agencia Estatal de Investigación (AEI) y el Fondo Europeo de Desarrollo Regional (FEDER).
How to cite: Manassero-Mas, A. M., Vázquez-Alonso, A. Assessing Critical Thinking Skills: A Diagnosis of Elementary Students (CT). *Revista Evaluar, 23*(2), 40-56. Retrieved from https://revistas.unc.edu.ar/index.php/revaluar
Participaron en la edición de este artículo: Agustina Mangieri, Gloria Rebeca Nieve, Juan Cruz Balverdi Nieto, Rita Hoyos, Florencia Ruiz, Jorge Bruera.

Introduction

Many institutions and experts worldwide support educating students for the skills of the 21st century to face the great challenges of today (European Union, 2014; Fullan & Scott, 2014; International Society for Technology Education, 2003; National Education Association, 2012; National Research Council, 2012; OECD, 2018; UNESCO, 2016). These skills include digital and cognitive skills; the latter usually distinguishes soft (psychosocial or interpersonal) and hard (higher-order cognitive) skills, which some authors summarize in the 4Cs or 6Cs (collaboration, communication, character, citizenship, creativity, and critical thinking [CT]). In sum, CT is a significant component of the skills for the 21st century, placing innovative demands on education (Almerich et al., 2020; Vincent-Lancrin et al., 2019).

From an educational perspective, CT teaching aligns with Piaget's pioneering studies (Piaget & Inhelder, 1997) and cognitive acceleration programs (Shayer & Adey, 2002) that have empirically demonstrated its significant impact on learning. In addition, the cognitive skills that make up CT are connected to the higher categories of Bloom's taxonomy (analyze, judge, and create), they are often called higher-order thinking skills. However, they also require the most basic skills, knowledge and understanding (Krathwohl, 2002). Nowadays, the mastery of CT skills is considered a key factor in achieving meaningful and deep learning skills that characterize educational excellence (Valenzuela, 2008). The meta-analysis of Hattie reports that the effect size of Piagetian programs on learning is very large ($d = 1.28$), and the impact of different CT skills (metacognitive strategies, creativity, problem-solving, etc.) is also high ($d > .40$) (Hattie, 2009).

From a labor perspective, most surveys show that CT is a primary and invariable requirement of future jobs (World Economic Forum, 2021) and a key factor for people's success in the information age (Tremblay et al., 2012). This labor requirement, coupled with the evolution of cognitive development, have driven most of the innovative teaching efforts of CT to be focused on higher education.

In sum, CT is a central objective of education, an important attribute of citizenship in a democratic society, and a decisive factor of an individual's professional success in the 21st century. These beneficial characteristics justify the attention placed on CT as a central variable of school learning. This study approaches this idea from a diagnostic evaluation perspective to address the lack of information about younger students' CT skills and aims to present this information from primary education and thus contribute to fill this gap.

Critical thinking

Research on CT has focused on 3 areas: conceptualization, teaching, and evaluation. However, the development of each area has been unequal (Saiz, 2017).

In the framework of cognitive psychology, CT is generally conceptualized as a type of thinking that masters multiple higher-order cognitive skills and various attitudinal dispositions, and is regulated by demanding quality standards (precision, solidity, coherence, relevance, adequacy, etc.) to overcome thinking's natural tendencies toward error, fallacy, and bias (egocentrism and socio-centrism). These skills, provisions, rules, and values inherent in CT provide a crucial basis for its evaluation (Bailin et al., 1999).

In contrast, the CT literature also shows a lack of consensus over a definition of CT due to the diversity of philosophical (e.g., Ennis, 2018; Facione, 1990; Paul & Nosich, 1993) and psychological (Halpern, 2003; Lai, 2011). approaches and concepts. A widely cited conceptualization of CT is the one proposed by Ennis (2018), who defines it as "reflective and reasonable thinking focused on deciding what to believe or do, along with its expanded development of the dispositions and skills involved in such decisions". To create some consensus among specialists, a panel of experts from the American Psychological Association (APA, 1990, p. 3) proposed a definition of CT as the "purposeful, self-regulatory judgment which results in interpretation, analysis, evaluation, and inference, as well as explanation of the evidential, conceptual, methodological, criteriological, or contextual considerations upon which that judgment is based". Many studies use this as a reference (APA, 1990; Facione, 1990).

As an alternative to this lack of conceptual consensus, some researchers choose to define CT by extension, that is, specifying its constitutive skills (Fisher, 2009). This approach is evident in the APA (1990) panel's definition, which mentions the skills of interpretation, analysis, evaluation, inference, judgment, and self-regulation. Ennis's (2018) expanded development also mentions decision-making. These examples or the extreme case of the so-called national plan for CT assessment (Paul & Nosich, 1993), which proposed a long list of 88 CT skills, evidence the lack of consensus on a definition of CT. However, some skills (e.g., analysis, reasoning, problem-solving, decision-making) and some dispositions (e.g., open-mindedness) can be considered predominant (Lai, 2011).

The different CT assessment instruments, by their functional and practical nature, tend to implicitly assume the extensive definition of CT, as each instrument usually specifies the skills it assesses. However, since the evaluation instruments are more specific than the definitions, here too, a lack of consensus is evident, highlighting the conceptual complexity of the CT construct. For example, the Cornell Critical Thinking Tests (CCTT) level X (Ennis et al., 2005a) assesses five dimensions (induction, deduction, observation, credibility, and assumptions), and the Critical Thinking Assessment test (HCTA) of Halpern (2007, 2010) assesses five skills (argument analysis, hypothesis-testing, probability and uncertainty, problem-solving, and verbal reasoning). In sum, the different skill terminologies, the unequal number of skills considered by each instrument (from 88 to 2), and the skills grouped into categories in some tools (those that offer a broader set of skills) are further examples of complexity, justifying the functionality of improving the organization of the CT field.

Some taxonomies of synthesis have been proposed to address this complexity and reduce the lack of consensus. For example, Dwyer et al. (2014) developed an integrated framework of educational objectives, cognitive processes (reflective judgment and self-regulation and meta-cognition functions), and CT skills (analysis, evaluation, and inference), including memory and comprehension as necessary processes to apply CT. Two recently developed taxonomies present a theoretical framework that organizes CT into four dimensions with significant coincidences between them. Manassero-Mas and Vázquez-Alonso (2019) proposed four basic dimensions of CT (creativity, reasoning and argumentation, complex processes, and evaluation and judgment), each containing multiple categories and subcategories (e.g., deductive, inductive, abductive, and statistical thinking; problem-solving and decision-making; assumptions, rules, dispositions). In a similar vein, Fisher (2021) also organized CT skills into four basic dimensions (interpretation, analysis, evaluation, and self-regulation), whose contents overlap broadly with the previous taxonomy.

In summary, the CT literature shows different conceptualizations among specialists, so to avoid misconceptions, we used the taxonomy of Manassero-Mas and Vázquez-Alonso (2019) as a general reference. According to the authors, the term CT is fundamental and consists of four dimensions, each containing multiple specific thinking skills and other associated concepts (dispositions and rules of attitude). The taxonomy also reflects most of the CT skills involved in most CT assessment instruments. Finally, despite the discrepancies presented, all the authors agree on the educational importance of CT.

The evaluation of critical thinking

CT can be taught and learned, multiple CT teaching programs with varied orientations and practices have attempted to teach CT for decades (Follmann et al., 2018; Saiz, 2017; Swartz et al., 2013). In addition, recommendation 12 of the APA expert statement (Facione, 1998) endorsed complementing the teaching of CT with its frequent and explicit evaluation, both diagnostic and summative (Recommendation 13), and using valid, reliable and equitable instruments which currently are obvious features in the construction of tests (Muñiz & Fonseca-Pedrero, 2019).

The need to evaluate is justified by Ennis (2018) on the following grounds: diagnosing the students' CT skills level, providing feedback on progress, motivating learning of CT, informing teachers about teaching methods, investigating CT, counseling on the choice of studies, and stimulating educational institutions to report their results. The evaluation of CT is a necessity and significant support for improving its teaching, but it requires the construction of appropriate evaluation instruments to achieve valid and reliable results.

The specialized literature offers numerous tests to assess CT and, although most focus on a few CT skills (e.g., Facione et al., 1998; Halpern, 2010; Rivas & Saiz, 2012; Watson & Glaser, 2002), others are broader (Madison, 2004). The analysis of the skills included in the CT assessment tests provides an overview of the CT skills synthesized in the CT taxonomies mentioned (Ennis & Chattin, 2018; Fisher, 2021; Manassero-Mas & Vázquez-Alonso, 2019). However, CT teaching programs that have proven their effects through empirical evaluation studies are the exception rather than the rule (Saiz, 2017). Lipman's (1982) Philosophy for Children program has been repeatedly evaluated (Colom et al., 2014), while others, such as thought-based learning (Swartz et al., 2013), have only been evaluated occasionally, and others, such as the reasoning program (Walton & Macagno, 2015) still lack evaluation.

The vast majority of CT assessment instruments target at adults and university students, and there are hardly any specific tests for young students, although the Cornell tests (X, Y, Z) are partially adaptable to different ages (Ennis & Millman, 2005a, 2005b), and other proposals require a consolidated development (Lopes et al., 2018). In addition, the review of Aktoprak and Hursen (2022) shows a great lack of research on CT in primary education and a predominance of qualitative research methodologies in the few existing works. Therefore, Gelerstein et al., 2016; Lai, 2011; Meng, 2016; Pérez-Morán et al., 2021; Sierra et al., 2010 propose guiding studies towards quantitative research methodologies that complement qualitative research methods and strengthening the evaluation of CT with reliable measurement tests. Also, the differences between men and women have also been rarely investigated, although the study by Sierra et al. shows no significant differences.

In sum, the educational development of CT has been unequal among the different educational levels (frequent in university and rare in the lower educational levels) and within the contents (teaching of CT has predominated, whereas reliable evaluation of CT is scarce, especially in younger students). These shortcomings justify this study's attention to the assessment of CT in young people, focused on specific skills appropriate and functional to their age range and contributing to drawing attention to the evaluation of CT in the early educational stages.

This study also builds on the development and evaluation of CT through the development of item banks on thinking skills for elementary school students (Manassero-Mas & Vázquez-Alonso, 2020a, 2020b). Based on the previous milestones and the application of psychometric recommendations to develop reliable tests (Fernández et al., 2010; Muñiz & Fonseca-Pedrero, 2019), a 48-item test that evaluates six thinking skills was validated. It is applied here to diagnose and highlight the thinking of primary school students. Its validity and reliability have been presented elsewhere (Manassero-Mas & Vázquez-Alonso, in press).

Consequently, the objectives of this study are: to quantitatively diagnose CT skills in 6th grade primary school children, present the normative data of the instrument, and compare the mastery of the skills in primary school children.

Method

Participants

The sample was comprised of 655 sixth-grade students (322 boys and 335 girls) with an average age of 11.16 years, who attended fourteen different schools in two Spanish communities (Catalonia, 42.6% and Balearic Islands, 57.4%), located in different towns (large, medium, small) of varied social contexts (upper, middle or lower class). Approximately half the participants studied in public schools (42.3%), and the other half (57.7%) in semi-private schools. All schools were selected for their favorable attitude towards critical thinking education. The students participated in this study in their own school groups, completing the thinking test as an assessment activity in the classroom under their teacher's direction.

Instrument

The test "Retos de Pensamiento" (RdP_EP6 [Thinking Challenges test]) applied in this study evaluates six CT skills: prediction and logical reasoning (reasoning dimension), comparison (creativity dimension), classification (evaluation dimension), and decision-making and problem-solving (complex processes dimension). These skills were agreed on with the schools participating in this study based on the skills adaptation to age and usual learning in sixth grade (EP6). The test items were designed using the criteria of readability, comprehensibility, balance on the cognitive demand of each item, the students' cognitive development, and the approach of facing a motivating and exciting challenge (Table 1).

Each test item was assigned to the skill most congruent with its content. For example, classification skill evaluates the ability to group or separate different elements according to their common or differential features. Prediction and comparison evaluate the ability to verify a logical conclusion through inductive reasoning or the creative contrast of several statements, respectively. Decision-making/problem-solving measures the ability to identify the best decisions/solutions in a particular situation, and logical reasoning evaluates simple (simple syllogism) and complex deductive ability (several pieces of information or conclusions are involved simultaneously).

The items propose a variety of scenarios and situations that communicate information by various means of representation (verbal, numerical, and figurative). One or more questions are asked, whose cognitive demand is adjusted to the students' skill and age expected average, posing authentic and motivating thinking challenges (see a sample in the appendix). The content of the items is independent of the curricula of the school subjects (for example, they do not propose numerical calculations), so the correct answer does not require previous school knowledge but only applying elemental skills to the information presented. Therefore, the applied test, RdP_EP6, is cultural-free; that is, its challenges are not mediated by social, familiar and academic knowledge as many thinking tests are. For example, the Science CT test requires knowledge of the primary science curriculum to answer correctly (Mapeala & Siew, 2015).

The RdP_EP6 response formats are mostly closed (four items require a short open answer) because this allows for a standardized, fast, valid, and reliable evaluation of each thinking skill and for developing diagnostic baselines to compare research, programs, and teaching methodologies. The reliability values of the six skill scales and the total test (Table 1) correspond to the empirical factors obtained by procedures described below (unweighted least squares [ULS], Manassero-Mas & Vázquez-Alonso, in press).

Procedures

The RdP_EP6 was applied to the participants in their class group by their teachers as a regulated ordinary evaluation to stimulate the students' effort and motivation. The application followed standardized guidelines using digital devices with no time limit for the answers (usually completed in a class period).

Correct answers received one point, incorrect answers received zero points, and no corrections were applied to random answers. The score of each skill is the sum of the correct answers in the items that comprise it, and the overall score is the sum of all the correct answers (estimation of the students' overall CT).

The validity of the content of the RdP_EP6 is based on the credibility of the specialized publications consulted for the original items (Ennis &

Table 1
Specifications of the test applied (RdP_EP6) in this study to evaluate thinking skills in the sixth grade of Primary Education EP6.

Thinking skills	Source	Type	Items	Reliability (ORION*)
Prediction (PREDIC)	Ennis & Millman. 2005a	Verbal	9	.86
Comparison (COMP)	Ennis & Millman. 2005a	Verbal	7	.74
Classification (CLAS)	Author elaboration **	Figurative	6	.91
Problem-solving (PROB)	Halpern (2010)	Verbal	6	.81
	Author elaboration **	Figurative	4	
Decision-Making (DECIS)	Author elaboration **	Mixed	9	.86
Logical reasoning (LOG-RA)	Ennis & Millman. 2005b	Verbal	7	.86
Total			48	(Alpha) .85

* Overall Reliability of fully-Informative prior Oblique N-Expected a Posteriori
** Translated and adapted from open materials of https://www.criticalthinking.com

Millman, 2005a, 2005b; Halpern, 2010), the items prepared by the authors (https://www.pensamientocritico.com), and the researchers' professional judgment for the consensual selection of the items. The criteria for item selection were the best fit between the item's content and the represented skill and between the item's cognitive demand and the students' cognitive level.

Analysis of results

The data of individual scores were processed with SPSS (25). The validity and reliability of the test were presented extensively (Manassero-Mas & Vázquez-Alonso, in press). They were calculated with the program Factor 12.01.02 (Ferrando & Lorenzo-Seva, 2017, 2018; Lorenzo-Seva & Ferrando, 2019), which applies a robust method of unweighted least squares (ULS) based on tetrachoric correlations, appropriate for dichotomous test scores, exploratory factor analysis (EFA), confirmatory factor analysis (CFA) extract factors with ULS and Promin rotation and evaluate reliability using various indices, such as ORION and Cronbach's alpha (Table 1).

The evaluation of the differences among groups calculates the degree of significance of the differences among groups (ANOVA). The effect size statistic (ES, d) measures the magnitude of the differences in standardized units of deviation, independent of the sample size and the test applied, unlike the degree of statistical significance (Funder & Ozer, 2019; Schäffer & Schwarz, 2019).

The central issue of the ES is to determine whether or not an effect is relevant, for which conventional reference points are usually applied, which vary according to the field of study (Cohen, 1988; Rosenthal, 1996; Ventura-León, 2018). Educational research often reports ESs lower than other disciplines; for example, the meta-analysis of Hattie (2009) adopts $d > .40$ as a reference of the practical relevance of the educational effect and $d > .60$ is considered large. In this study, practical educational relevance was attributed to $d > .20$ because the probability is usually already statistically significant, and the following references were adopted for ES: Small ($d < .20$), medium (.20 - .30), moderate (.30 - .50), and large (> .50).

Results

The overall results of the 48 items that make up the RdP_EP6 are summarized in Table 2. The global average of the 48 items is .492, indicating that the test has a medium difficulty rate, very close to 50% of correct answers. In addition, there are six very difficult items (hit rate less than .30), and five very easy items (hit rate greater than .70), so 81% of the items have medium difficulty indexes included in the central range (.30 - .70).

Table 3 presents the descriptive results of the scores in the six thinking skills evaluated by the RdP_EP 6, obtained by adding the correct responses to the items that are part of each skill. As the number of items for each skill is different, the means obtained are not directly comparable. However, taking as a reference the central point of the scale of each skill, the results show that the prediction, classification, and problem-solving scales have means above their midpoint, whereas the comparison, decision-making, and logical reasoning scales obtain means below their midpoint. Hence, the former skills obtain overall hits above 50% (the easiest), whereas the latter ones obtain success rates below 50% (more difficult for the students).

Table 2
Descriptive statistics of the 48 items of the RdP_EP6 test (N = 655).

Variables		Mean	SD	Standard Error	95% Confidence interval of the mean	
					Lower limit	Upper limit
V1	PREDIC1	.62	.48	.02	.59	.66
V2	PREDIC2	.43	.50	.02	.39	.47
V3	PREDIC3	.50	.50	.02	.46	.54
V4	PREDIC4	.40	.49	.02	.36	.44
V5	PREDIC5	.79	.41	.02	.76	.82
V6	PREDIC6	.74	.44	.02	.70	.77
V7	PREDIC7	.71	.45	.02	.67	.74
V8	PREDIC8	.38	.49	.02	.34	.42
V9	PREDIC9	.64	.48	.02	.60	.67
V10	COMPA1	.44	.50	.02	.40	.48
V11	COMPA2	.56	.50	.02	.53	.60
V12	COMPA3	.43	.50	.02	.39	.47
V13	COMPA4	.52	.50	.02	.48	.56
V14	COMPA5	.50	.50	.02	.46	.54
V15	COMPA6	.50	.50	.02	.46	.54
V16	COMPA7	.36	.48	.02	.32	.39
V17	CLASIF1	.64	.48	.02	.60	.67
V18	CLASIF2	.55	.50	.02	.51	.59
V19	CLASIF3	.56	.50	.02	.52	.60
V20	CLASIF4	.65	.48	.02	.62	.69
V21	CLASIF5	.66	.47	.02	.63	.70
V22	CLASIF6	.64	.48	.02	.60	.68
V23	PROBL1	.65	.48	.02	.61	.69
V24	PROBL2	.61	.49	.02	.57	.65
V25	PROBL3	.57	.49	.02	.53	.61
V26	PROBL4	.32	.47	.02	.29	.36
V27	PROBL5	.73	.45	.02	.69	.76
V28	PROBL6	.73	.44	.02	.70	.77
V29	DECIS1	.35	.48	.02	.31	.39
V30	DECIS2	.28	.45	.02	.25	.32
V31	DECIS3	.23	.46	.02	.26	.33
V32	DECIS4	.33	.47	.02	.29	.36
V33	DECIS5	.43	.49	.02	.39	.46
V34	DECIS6	.19	.39	.02	.16	.22
V35	DECIS7	.15	.35	.02	.118	.17
V36	DECIS8	.65	.48	.02	.61	.68
V37	DECIS9	.46	.50	.02	.43	.50
V38	PROBL9	.21	.41	.02	.18	.24
V39	PROBL10	.43	.49	.02	.39	.46

Variables		Mean	SD	Standard Error	95% Confidence interval of the mean	
					Lower limit	Upper limit
V40	PROBL11	.54	.50	.02	.50	.57
V41	PROBL12	.38	.49	.02	.34	.41
V42	LOGIC1	.54	.50	.02	.50	.58
V43	LOGIC2	.55	.50	.02	.52	.59
V44	LOGIC3	.30	.47	.02	.27	.34
V45	LOGIC4	.59	.50	.02	.55	.63
V46	LOGIC5	.24	.42	.02	.20	.27
V47	LOGIC6	.57	.49	.02	.54	.61
V48	LOGIC7	.33	.47	.02	.29	.36

Table 3
Descriptive statistical results of the six thinking skills and the total score evaluated with the RdP_EP6 test.

Skills	Items	Mean	SD	Standard Error	95% Confidence interval of the mean		Minimum	Maximum
					Lower limit	Upper limit		
PREDICTION9	9	5.21	1.92	.07	5.06	5.35	0	9
COMPARISON7	7	3.31	1.42	.05	3.20	3.42	0	7
CLASSIFICATION6	6	3.70	1.89	.07	3.56	3.85	0	6
PROBLEMS10	10	5.17	2.14	.08	5.00	5.33	0	10
DECISIONS9	9	3.12	1.85	.07	2.98	3.27	0	9
LOGIC7	7	3.12	1.63	.06	3.00	3.25	0	7
TOTAL48	48	23.64	6.86	.27	23.11	24.16	10	45

The test's average total score (23.64) is close to 24, which marks the central point of the overall score, reflecting the intermediate global difficulty, close to 50%, in direct scores of the complete test. Also, the table indicates that the responses in all the skills reach the minimum (0) and maximum scores, which means that some students did not answer any item of the skill correctly, but also that some students answered all the items of each skill correctly. Regarding the global test, the minimum score achieved is 10 correct answers, and the maximum score is 45 correct answers, much closer to the possible maximum score (48) than the minimum score (10) regarding the possible minimum score (0 points).

Scale and normative data of the test

Table 4 presents the frequency distribution of the total RdP_EP6_48 scores obtained by the sample of students. The range extends from the minimum score of 10 correct answers to the maximum score of 45 correct answers. The mean score is 23.64 (Table 2), the median is 22, and the mode is 19.

The standardization of these scores in quartiles shows an asymmetric and distorted curve towards the highest scores because the highest quartile includes from score 28 to the maximum score of 45 (half the range of the scores obtained). This range is practically identical to the range of scores in the lower three quartiles (from the minimum score of 10 to score 27). Similarly, the distribution

Table 4
Standardized distribution of total RdP_EP6 test scores in the sample of primary school students.

Points	N	%	Percentiles	Quartiles
10	5	0.8		
11	9	2.1		
12	7	3.2		
13	9	4.6		
14	20	7.6	10	
15	28	11.9		
16	24	15.6		
17	29	20	20	
18	26	24		25
19	48	31.3	30	
20	37	36.9		
21	40	43.1	40	
22	37	48.7	50	50
23	33	53.7		
24	30	58.3	60	
25	23	61.8		
26	33	66.9		
27	25	70.7	70	
28	17	73.3		75
29	27	77.4		
30	27	81.5	80	
31	28	85.8		
32	25	89.6	90	
33	11	91.3		
34	12	93.1		
35	7	94.2		
36	12	96		
37	10	97.6		
38	3	98.0		
39	6	98.9		
40	3	99.4		
41	2	99.7		
44	1	99.8		
45	1	100		
Total	655			

of scores is strongly concentrated in the central percentile sections (between the 20th and 80th percentiles), which practically encompass one, two, or three different scores, whereas the lowest percentile section (10) comprises five different scores (between 10 and 14), and the highest percentile section (90) comprises 13 different scores (between 32 and 45).

The distribution of the scores on the six scales of the CT skills of the RdP_EP6 obtained by the sample of students is presented in Table 5. The range of the six scales is different, so the maximum scores vary according to the skill, from the shortest range of the classification skill (6) to the longest range of the problem-solving skill (10).

Gender differences in thinking skills

To evaluate the gender differences in thinking skills, we compared the scores obtained by the groups of boys and girls in all the variables of the RdP_EP6 considered in this study, which meet the conditions of normality, equality of variances, and sample similarity. The relevance of the differences between the two groups was measured with two statistics: the degree of significance of the differences (through ANOVA) and the ES of the differences (through Cohen's formula, as the two groups are similar in size). The ES was computed subtracting the girls' average to the boys'mean. Thus, positive differences indicate the boys' higher score, and negative differences indicate the girls' higher score.

Table 6 presents the results of the means and standard deviations for each of the 48 items that make up the test of the two compared groups of boys and girls, the two statistics assessing the differences, the degree of significance of the differences (p) and the ES of the differences (d), ordered from highest to lowest according to the ES.

Table 5
Distribution of scores on the six CT skill scales of the RdP_EP6 test.

Points	PREDIC9		COMPA7		CLASIF6		PROBL10		DECIS9		LOGIC7	
	N	%	N	%	N	%	N	%	N	%	N	%
0	5	0.8	7	1.1	29	4.4	4	0.6	31	4.7	33	5
1	18	3.5	54	9.3	72	15.4	14	2.7	102	20.3	78	16.9
2	43	10.1	142	31	103	31.1	49	10.2	143	42.1	140	38.3
3	66	20.2	167	56.5	94	45.5	91	24.1	131	62.1	126	57.6
4	79	32.2	140	77.9	88	58.9	109	40.8	93	76.3	135	78.2
5	134	52.7	102	93.4	101	74.4	118	58.8	80	88.5	97	93
6	135	73.3	40	99.5	168	100	74	70.1	44	95.3	37	98.6
7	106	89.5	3	100			88	83.5	21	98.5	9	100
8	51	97.3					70	94.2	7	99.5		
9	18	100					27	98.3	3	100		
10							11	100				
Total	655		655		655		655		655		655	

Table 6
Descriptive statistics of the 48 items of the RdP_EP6 test for boys and girls, the degree of significance and the effect size of the group differences ordered by effect size.

Variables		Boys		Girls		P-Sig.	Effect size
		Mean	SD	Mean	SD		
V38	PROBL9	.25	.44	.17	.38	.009	.195
V35	DECIS7	.17	.37	.13	.33	.144	.114
V13	COMPA4	.55	.50	.50	.50	.189	.100
V31	DECIS3	.31	.46	.28	.45	.419	.066
V4	PREDIC4	.41	.49	.39	.49	.750	.041
V12	COMPA3	.44	.50	.42	.49	.725	.040
V46	LOGIC5	.24	.43	.23	.42	.611	.024
V44	LOGIC3	.31	.46	.30	.46	.827	.022
V7	PREDIC7	.71	.45	.71	.46	.957	.000
V3	PREDIC3	.49	.50	.50	.50	.784	-.020
V39	PROBL10	.42	.49	.43	.50	.837	-.020
V23	PROBL1	.64	.48	.65	.48	.789	-.021
V26	PROBL4	.32	.47	.33	.47	.731	-.021
V27	PROBL5	.72	.45	.73	.44	.786	-.022
V2	PREDIC2	.42	.49	.44	.50	.452	-.040
V15	COMPA6	.49	.50	.51	.50	.726	-.040
V41	PROBL12	.37	.48	.39	.49	.554	-.041
V5	PREDIC5	.78	.42	.80	.40	.435	-.049
V48	LOGIC7	.31	.46	.34	.47	.449	-.065
V28	PROBL6	.72	.45	.75	.43	.290	-.068
V42	LOGIC1	.52	.50	.56	.50	.277	-.080

Variables		Boys		Girls		P-Sig.	Effect size
		Mean	SD	Mean	SD		
V10	COMPA1	.42	.49	.46	.50	.258	-.081
V1	PREDIC1	.60	.49	.64	.48	.238	-.082
V22	CLASIF6	.62	.49	.66	.47	.210	-.083
V30	DECIS2	.26	.44	.30	.46	.269	-.089
V43	LOGIC2	.53	.50	.58	.49	.142	-.101
V17	CLASIF1	.61	.49	.66	.47	.182	-.104
V36	DECIS8	.62	.49	.67	.47	.158	-.104
V6	PREDIC6	.71	.45	.76	.43	.185	-.114
V14	COMPA5	.47	.50	.53	.50	.172	-.120
V25	PROBL3	.54	.50	.60	.49	.091	-.121
V8	PREDIC8	.35	.48	.41	.49	.121	-.124
V29	DECIS1	.32	.47	.38	.49	.125	-.125
V21	CLASIF5	.63	.48	.69	.46	.136	-.128
V34	DECIS6	.16	.37	.21	.41	.152	-.128
V40	PROBL11	.50	.50	.57	.50	.101	-.140
V33	DECIS5	.39	.49	.46	.50	.104	-.141
V45	LOGIC4	.55	.50	.62	.49	.055	-.141
V16	COMPA7	.32	.47	.39	.49	.077	-.146
V20	CLASIF4	.62	.49	.69	.46	.047	-.147
V19	CLASIF3	.52	.50	.60	.49	.030	-.162
V9	PREDIC9	.59	.49	.68	.47	.013	-.188
V18	CLASIF2	.50	.50	.60	.49	.013	-.202
V24	PROBL2	.56	.50	.66	.48	.013	-.204
V32	DECIS4	.28	.45	.38	.49	.008	-.213
V37	DECIS9	.41	.49	.52	.50	.004	-.222
V47	LOGIC6	.52	.50	.63	.48	.005	-.224
V11	COMPA2	.50	.50	.62	.49	.002	-.242

The main finding comparing boys and girls is that most of the differences obtained in all the 48 items show that girls score higher than boys in 39 items (negative ES), and boys score higher in only eight of the remaining items (positive ES). This result indicates that girls in the sixth grade of primary education have, on average, better CT skills than boys.

The second finding in Table 6 is that the differences between boys are low and mostly non-significant. Indeed, all the differences between boys and girls calculated through the ES are less than .30, and among the highest that obtain negative values favoring the girls, only six items exceed the value .20. Similarly, only nine items reach a significance level of $p < .05$ (of which four items reach $p < .01$). In sum, the significance and ES of the differences between boys and girls are small.

The results obtained for the gender differences between primary school boys and girls in the six skill variables and the total score of the questionnaires confirm and reinforce the patterns and trends found for the 48 items of the test, given the additive nature of the skill scales (Table 7).

All the differences in the six skills and the total score favor the girls (negative), which shows the overwhelming dominance of girls in almost all the items. Although most scores achieve statistically significant differences, the ES of the differences remains low. The largest gender difference is in the total score, and the smallest is in problem-solving skills.

Table 7
Descriptive statistics of the six CT skills and the total score of the RdP_EP6 test (means and standard deviations) in the group of boys and girls, with the degree of significance and the effect size of the group differences.

Skills	Boys		Girls		P-Sig.	Effect size
	Mean	SD	Mean	SD		
PREDICTION9	5.06	1.91	5.35	1.93	.051	-.151
COMPARISON7	3.19	1.41	3.43	1.42	.035	-.170
CLASSIFICATION6	3.49	1.89	3.90	1.87	.006	-.218
PROBLEMS10	5.04	2.18	5.28	2.09	.151	-.112
DECISIONS9	2.92	1.80	3.32	1.88	.006	-.217
LOGIC7	2.98	1.67	3.26	1.58	.025	-.172
TOTAL48	22.69	6.89	24.54	6.72	.001	-.272

Discussion and conclusions

The main objective of this study is to diagnose the level of CT skills in a large sample of sixth-grade students of Primary Education (11 years) through the RdP_EP6 test, which evaluates six CT skills (prediction, comparison, classification, problem-solving, decision-making, and logical reasoning). The results indicate that the students reach an intermediate level of mastery of CT skills (about 50% of correct answers in the global test) on the cognitive demands of the test items, a first reference of mastery for this sample and this test. Concerning the relative mastery of the different skills assessed in the test, the students show a greater relative mastery of prediction, classification, and problem-solving skills (scores above the midpoint of each measurement scale), whereas comparison, decision-making, and logical reasoning skills have relatively lower scores (below the midpoint of each measurement scale). These results are complemented with the psychometric evaluation of the test and the scales of the six skills, which can serve as a global reference framework for the expansion of the test's standardization with different samples from other contexts and places, contributing to the development of a valid and reliable test (Manassero-Mas & Vázquez-Alonso, in press).

CT studies in primary education are few and qualitative (Gelerstein et al., 2016; Lai, 2011; Meng, 2016; Pérez-Morán et al., 2021). In addition, there is a lack of specific tests to evaluate CT in youngsters and there are even fewer studies evaluating skills which do not assess students' real mastery of CT skills. For example, Lopes et al. (2018) developed a qualitative test for students from 12 to 19 years old, and Pérez-Moran et al. (2021) did so quantitatively, but they did not value the real mastery of the students' performance. In short, there is a lack of quantitative studies that can serve as a reference to assess the domain of CT reflected in the scores of the skills evaluated in primary education. This prevents contrasting the scope and value of the results obtained in the sample of this study with other equivalent samples evaluated with different instruments. Thus, these results are pioneer in serving as a precedent and diagnostic reference for subsequent studies and they contribute to filling the gaps, although the test's valuation is pending future confirmation.

The most notable finding is the girls' higher level of CT skills in most test items, the six skills and the total CT score. The differences in favor of girls are statistically significant in comparison, classification, decision-making, logical reasoning and, of course, the total CT score. In sum, although the magnitudes of the differences are not large, the statistically significant superiority of girls over boys in CT skills constitute a consistent and solid trend. This supports girls' better performance instead of ratifying a hypothesis of similarity of the two groups in primary school students (Sierra et al., 2010) or the differences in older students, obtained with statistics inappropriate to the group size (Lopes et al., 2018).

Girls' better CT mastery suggests two interesting facts. The first refers to the justification and explanation of this differential result because if boys and girls have mostly attended the same school together, in the same classes, and with the same teachers, there is no evidence to attribute the differences to cultural or educational variables. Thus, the explanatory parameters could be within the framework of the evolutionary differences between boys and girls.

An additional interesting issue is related to the hypothesis of similarity of men and women presented in the literature of differential psychology, where it is still considered that spatial mental rotation is the only capacity that presents large empirical differences favoring men, controlling for the educational and cultural background (Jäncke et al., 2018). Item V2638-PROBL9 of the RdP_EP6 makes a cognitive demand that involves imagining the rotation of a cube to give the correct answer, and its result of gender differences (Table 6) presents the greatest magnitude of the differences favoring boys (*d* = 0.195), consistent with differential psychology's prediction about spatial rotation. This result confirms differential psychology research on spatial rotation and further supports the validity and reliability of the RdP_EP6.

References

Aktoprak, A., & Hursen, C. (2022). A bibliometric and content analysis of critical thinking in primary education. *Thinking Skills and Creativity, 44*, 101029. https://doi.org/10.1016/J.TSC.2022.101029

Almerich, G., Suárez-Rodríguez, J., Díaz-García, I., & Orellana, N. (2020). Estructura de las competencias del siglo XXI en alumnado del ámbito educativo. Factores personales influyentes [Structure of 21st century competences in students the sphere of education. Influential personal factors]. *Educación XX1, 23*(1), 45-74. https://doi.org/10.5944/educXX1.23853

American Psychological Association APA (1990). *Critical thinking: A statement of expert consensus for purposes of educational assessment and instruction (The Delphi Report)*. https://philarchive.org/archive/faccta

Bailin, S., Case, R., Coombs, J. R., & Daniels, L. B. (1999). Conceptualizing critical thinking. *Journal of Curriculum Studies, 31*(3), 285-302. https://doi.org/10.1080/002202799183133

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*, 2nd ed. Academic Press.

Colom, R., García Moriyón, F., Magro, C. & Morilla, E. (2014). The long-term impact of philosophy for children: A longitudinal study (Preliminary results). *Analytic Teaching and Philosophical Praxis, 35*(1), 50-56. https://journal.viterbo.edu/index.php/atpp/article/view/1129

Dwyer, C. P., Hogan, M. J., & Stewart, I. (2014). An integrated critical thinking framework for the 21st century. *Thinking Skills and Creativity, 12*, 43-52. https://doi.org/10.1016/J.TSC.2013.12.004

Ennis, R. H. (2018). Critical thinking across the curriculum: A vision. *Topoi, 37*, 165-184. https://doi.org/10.1007/

s11245-016-9401-4

Ennis, R. H. & Chattin, G. S. (2018). An annotated list of critical thinking tests. http://criticalthinking.net/wp-content/uploads/2018/01/An-Annotated-List-of-English-Language-Critical-Thinking-Tests.pdf

Ennis, R. H., & Millman, J. (2005a). *Cornell Critical Thinking Test Level X.* The Critical Thinking Company.

Ennis, R. H., & Millman, J. (2005b). *Cornell Critical Thinking Test Level Z.* The Critical Thinking Company.

European Union (2014). *Key competence development in school education in Europe. KeyCoNet's review of the literature: A summary*. European Schoolnet. http://keyconet.eun.org

Facione, P. A. (1990). *Critical thinking: A statement of expert consensus for purposes of educational assessment and instruction. Research findings and recommendations*. American Philosophical Association. https://eric.ed.gov/?id=ED315423

Facione, P. A. (1998). *Insight assessment*. www.insightassessment.com

Facione, P. A., Facione, N. C., Blohm, S.W., Howard, K., & Giancarlo, C. A. F. (1998). *California Critical Thinking Skills Test: Manual (Revised)*. California Academic Press.

Fernández, A., Pérez, E., Alderete, A. M., Richaud, M. C., & Fernández Liporace, M. (2010). ¿Construir o adaptar tests psicológicos? Diferentes respuestas a una cuestión controvertida. *Revista Evaluar, 10*(1). https://doi.org/10.35670/1667-4545.v10.n1.459

Ferrando, P. J., & Lorenzo-Seva, U. (2017). Program FACTOR at 10: Origins, development and future directions. *Psicothema, 29*, 236-240. https://doi.org/10.7334/psicothema2016.304

Ferrando, P. J., & Lorenzo-Seva U. (2018). Assessing the quality and appropriateness of factor solutions and factor score estimates in exploratory item factor analysis. *Educational and Psychological Measurement, 78*, 762-780. https://doi.org/10.1177/0013164417719308

Fisher, A. (2009). Critical thinking. An introduction. Cambridge University Press.

Fisher, A. (2021). On what critical thinking is. In J. A. Blair (Ed.), *Studies in critical thinking* (2nd ed., pp. 7-26). University of Windsor. https://doi.org/10.22329/wsia.08.2019

Follmann, D., Mattos, K. R. C., & Güllich, R. I. da C. (2018). Teaching strategies of sciences and the promotion of critical thinking in Portugal. *Tecné, Episteme y Didaxis* (Extraordinario, Octavo Congreso Internacional de formación de Profesores de Ciencias para la Construcción de Sociedades Sustentables). https://revistas.pedagogica.edu.co/index.php/ESD/article/view/8789

Fullan, M., & Scott, G. (2014). *Education PLUS*. Collaborative Impact SCT. https://michaelfullan.ca

Funder, D. C., & Ozer, D. J. (2019). Evaluating effect size in psychological research: Sense and nonsense. *Advances in Methods and Practices in Psychological Science, 2*(2), 156-168. https://doi.org/10.1177/2515245919847202

Gelerstein, D., del Río, R., Nussbaum, M., Chiuminatto, P., & López, X. (2016). Designing and implementing a test for measuring critical thinking in primary school. *Thinking Skills and Creativity, 20*, 40-49. https://doi.org/10.1016/J.TSC.2016.02.002

Halpern , D. F. (2003). *Thought and knowledge: An introduction to critical thinking*. Laurence Erlbaum Associates.

Halpern, D. F. (2007). *Halpern Critical Thinking Assessment using everyday situations: Background and scoring standards*. Claremont McKenna College.

Halpern , D. F. (2010). *Manual Halpern Critical Thinking Assessment*. Schuhfried GmbH.

Hattie, J. (2009). *Visible learning: A synthesis of over 800 meta-analyses relating to achievement*. Routledge.

International Society for Technology Education. (2003). *National educational technology standards for teachers: Preparing teachers to use technology*. ISTE.

Jäncke, L. (2018). Sex/gender differences in cognition, neurophysiology, and neuroanatomy. *F1000Research, 7*,805. https://doi.org/10.12688/

f1000research.13917.1

Krathwohl, D. (2002). A revision of Bloom's taxonomy: An Overview. *Theory into Practice, 41*, 212-218. https://doi.org/10.1207/s15430421tip4104_2

Lai, E. R. (2011). Critical thinking: A literature review. *Pearson's research reports, 6*, 1-49.

Lipman, M. (1982). Philosophy for Children. *Thinking: The Journal of Philosophy for Children, 3*(3/4), 35-44. https://doi.org/10.5840/thinking1982339

Lopes, J., Silva, H., & Morais, E. (2018). Teste de pensamento crítico para estudantes dos ensinos básico e secundário [Critical thinking test for elementarý and secondary students]. *Revista de Estudios e Investigación en Psicología y Educación, 5*(2), 82-91. https://doi.org/10.17979/reipe.2018.5.2.3339

Lorenzo-Seva, U., & Ferrando, P. J. (2019). Robust promin: A method for diagonally weighted factor rotation. *LIBERABIT, Revista Peruana de Psicología, 25*, 99-106. https://doi.org/10.24265/liberabit.2019.v25n1.08

Madison, J. (2004). *James Madison critical thinking course.* The Critical Thinking Co.

Manassero-Mas, M. A., & Vázquez-Alonso, A. (2019). Taxonomía de las destrezas de pensamiento: una herramienta clave para la alfabetización científica [Taxonomy of thinking skills: A key tool for scientific literacy]. In M. D. Maciel & E. Albrecht (Org.), *Ciência, Tecnologia & Sociedade: Ensino, Pesquisa e Formação* (pp. 17-38). UNICSUL.

Manassero-Mas, M. A., & Vázquez-Alonso, A. (2020a). Evaluación de destrezas de pensamiento crítico: Validación de instrumentos libres de cultura [Assessment of critical thinking skills: Validation of free-culture tools]. *Tecné, Episteme y Didaxis, 47*, 15-32. https://doi.org/10.17227/ted.num47-9801

Manassero-Mas, M. A., & Vázquez-Alonso, A. (2020b). Las destrezas de pensamiento y las calificaciones escolares en educación secundaria: Validación de un instrumento de evaluación libre de cultura [Thinking skills and school grades in secondary education: Validation of a culture-free assessment instrument]. *Tecné, Episteme y Didaxis, 48*, 33-54. https://doi.org/10.17227/ted.num48-12375

Manassero-Mas, M. A., & Vázquez-Alonso, A. (in press). Assessment of critical thinking skills in primary education: Validation of Challenges of Thinking Test. *Thinking Skills and Creativity*.

Mapeala, R., & Siew, N. M. (2015). The development and validation of a test of science critical thinking for fifth graders. *SpringerPlus, 4*(1). https://doi.org/10.1186/s40064-015-1535-0

Meng, K.H. (2016). Infusion of critical thinking across the English language curriculum: A multiple case study of primary school in-service expert teachers in Singapore. Ph.D. Thesis, University of Western Australia, Perth, Australia.

Muñiz, J., & Fonseca-Pedrero, E. (2019). Diez pasos para la construcción de un test [Ten steps for test development]. *Psicothema, 31*(1), 7-16. https://doi.org/10.7334/psicothema2018.291

National Education Association. (2012). *Preparing 21st-century students for a global society: An educator's guide to the "four Cs"*. National Education Association.

National Research Council. (2012). *Education for life and work: Developing transferable knowledge and skills in the 21st century*. The National Academies Press. https://doi.org/10.17226/13398

Organization for Economic Co-operation and Development. (2018). The future of education and skills. *Education 2030*. OECD Publishing. https://www.oecd.org/education/2030-project/contact

Paul, R., & Nosich, G. M. (1993). A model for the national assessment of higher order thinking. In R. Paul (Ed.), *Critical thinking: What every student needs to survive in a rapidly changing world* (pp. 78-123). Foundation for Critical Thinking.

Pérez-Morán, G., Bazalar-Palacios, J., & Arhuis-Inca, W. (2021). Diagnóstico del pensamiento crítico de estudiantes de educación primaria de Chimbote, Perú [Diagnosis of critical thinking of elementary school students in Chimbote, Peru]. *Revista*

*Electrónica Educare, 25(*1), 289-299. https://dx.doi.org/10.15359/ree.25-1.15

Piaget, J., & Inhelder, B. (1997). *Psicología del niño* [The psychology of the child]. Morata.

Rivas, S. F., & Saiz, C. (2012). Validación y propiedades psicométricas de la prueba de pensamiento crítico PENCRISAL. *Revista Electrónica de Metodología Aplicada, 17*(1), 18-34. https://reunido.uniovi.es/index.php/Rema/index

Rosenthal, J. A. (1996). Qualitative descriptors of strength of association and effect size. *Journal of Social Service Research, 21*(4), 37-59. https://doi.org/10.1300/J079v21n04_02

Saiz-Sánchez, C. (2017). *Pensamiento crítico y cambio* . Pirámide.

Schäfer, T., & Schwarz, M. A. (2019). The meaningfulness of effect sizes in psychological research: Differences between sub-disciplines and the impact of potential biases. Frontiers in Psychology, 10. https://doi.org/10.3389/fpsyg.2019.00813

Shayer, M., & Adey, P. S. (Eds.) (2002). *Learning intelligence: Cognitive acceleration across the curriculum from 5 to 15 years*. Open University Press.

Sierra-Paz, J., Carpintero-Molina, E., & Pérez-Sánchez, L. (2010). Pensamiento crítico y capacidad intelectual. *Faísca Revista de divulgación científica sobre las altas capacidades intelectuales, 15*(17), 98-110. https://www.revistafaisca.es

Swartz, R. J., Costa, A. L., Beyer, B. K., Reagan R., & Kallick, B. (2013). *El aprendizaje basado en el pensamiento. Cómo desarrollar en los alumnos las competencias del siglo XXI* [Thinking-based learning. Promoting quality student achievement in the 21st Century]. Ediciones SM.

Tremblay, K., Lalancette, D., & Roseveare, D. (2012). AHELO. *Assessment of higher education learning outcomes. Volume 1 - Design and implementation (Feasibility Study Report)*. OECD. http://www.oecd.org/education/skills-beyond-school

UNESCO. (2016). *Education 2030: Incheon declaration and framework for action for the implementation of sustainable development goal 4: Ensure inclusive and equitable quality education and promote lifelong learning opportunities for all*. https://unesdoc.unesco.org

Valenzuela, J. (2008). Habilidades de pensamiento y aprendizaje profundo. *Revista Iberoamericana de Educación, 46*(7), 1-9. https://doi.org/10.35362/rie4671914

Ventura-León, J. (2018). Otras formas de entender la d de Cohen [Other ways of understanding Cohen's d]. *Revista Evaluar, 18*(3). https://doi.org/10.35670/1667-4545.v18.n3.22305

Vincent-Lancrin, S., González-Sancho, C., Bouckaert, M., de Luca, F., Fernández-Barrerra, M., Jacotin, G., Urgel, J., & Vidal, Q. (2019). *Fostering students' creativity and critical thinking*. OECD. https://doi.org/10.1787/62212c37-en

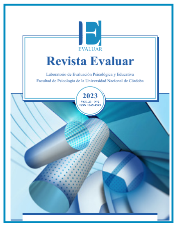
Walton, D., & Macagno, F. (2015). A classification system for argumentation schemes. *Argument and Computation, 6*(3), 214-249. https://www.tandfonline.com/loi/tarc20

Watson, G., & Glaser, E. M. (2002). *Watson-Glaser Critical Thinking Appraisal-II Form E*. Pearson.

World Economic Forum (2021). *These are the top 10 job skills of tomorrow - and how long it takes to learn them*.

2023, Vol. 23, No. 2
ISSN 1667-4545
Recuperado de https://revistas.unc.edu.ar/index.php/revaluar

Laboratorio de Evaluación Psicológica y Educativa
Facultad de Psicología - Universidad Nacional de Córdoba



Validation of the White Bear Suppression Inventory for the Mexican population

Validación del Inventario de Supresión del Oso Blanco en Población Mexicana

Julia Gallegos-Guajardo [1] , Gabriela Sánchez-Jáuregui [1] , Liz Alicea-Lozada [2] , Johanna Hidalgo [2] , Vania Luna-Huguenin [1] , Brian Fisak * [2]

1 - *School of Psychology, Universidad de Monterrey. Av. Ignacio Morones Prieto 4500 poniente. Col. Jesús M. Garza San Pedro Garza García Nuevo León, México C. P. 66238.*
2 - *Department of Psychology, University of Central Florida, Sanford Lake Mary Regional Campus, 100 Weldon Boulevard, Partnership Center, Sanford, FL, 32773.*

Recibido: 23/02/2023 **Revisado:** 09/04/2023 **Aceptado:** 19/04/2023

Abstract

Thought suppression is the general tendency to suppress unwanted, negative thoughts and the inability to ignore unwanted intrusive thoughts. The purpose of this study was to develop a translated version of the White Bear Suppression Inventory (WBSI) that is reliable and valid in a Mexican sample. A back-to-back translation of the WBSI was made, and the WBSI- Mexican Version (WBSI-MV) was administered to a sample of 346 undergraduate students enrolled at university in northern Mexico. Support was found for a two-factor solution, with factors labeled intrusion and suppression, which is consistent with previous research. Furthermore, support was found for the validity of the WBSI-MV, as both subscales were considered to be significantly and positively associated with measures of obsessive-compulsive disorder symptoms and thought-action fusion. Overall, the WBSI-MV proved to be a valid measure that can be used to assess suppressive and intrusive thoughts in the Mexican population.

Keywords: *thought suppression, White Bear Suppression Inventory, intrusive thoughts, obsessive-compulsive disorder, Mexican population*

Resumen

El Inventario de Supresión del Oso Blanco (WBSI) se encuentra entre los instrumentos de autoinforme más utilizados para evaluar la supresión del pensamiento, la tendencia general a suprimir los pensamientos negativos no deseados y la incapacidad de ignorarlos. El propósito del presente estudio fue desarrollar una versión traducida del WBSI que sea confiable y válida en una muestra mexicana. Se realizó una traducción y se administró a 346 estudiantes universitarios en el norte de México. Los resultados mostraron una solución de dos factores, etiquetados como intrusión y supresión, lo cual es consistente con investigaciones previas. Además, se apoyó la validez de la versión mexicana (WBSI-MV), ya que ambas subescalas estaban asociadas de manera significativa y positiva con medidas de síntomas de trastorno obsesivo compulsivo y fusión de pensamiento-acción. En general, se mostró que el WBSI-MV es un instrumento válido que puede utilizarse para evaluar los pensamientos supresivos e intrusivos en la población mexicana.

Palabras clave: *supresión del pensamiento, Inventario de Supresión del Oso Blanco, población mexicana, pensamientos intrusivos, trastorno obsesivo compulsivo*

***Correspondence to:** Brian Fisak, Department of Psychology, University of Central Florida, Sanford Lake Mary Regional Campus, 100 Weldon Boulevard, Partnership Center, Sanford, FL, 32773. Phone number +1 407 708 247. Email address: brian.fisak@ucf.edu

How to cite: Gallegos-Guajardo, J., Sánchez-Jáuregui, G., Alicea-Lozada, L., Hidalgo, J., Luna-Huguenin, V., Fisak, B, Validation of the White Bear Suppression Inventory for the Mexican population (WBSI-MV). *Revista Evaluar, 23*(2), 57-66. https://revistas.unc.edu.ar/index.php/revaluar

Participaron en la edición de este artículo: María Eugenia Maiorana, Juan Cruz Balverdi Nieto, Pablo Carpintero, Florencia Ruiz, Rodrigo Maderna, Jorge Bruera.

Introduction

Thought suppression involves keeping thoughts out of the mind or consciousness (Wegner & Zanakos, 1994). Interestingly, thought suppression may have an ironic or paradoxical effect, as attempts to suppress thoughts may actually increase the occurrence of these thoughts (Pica et al., 2015; Wegner & Zanakos, 1994). In particular, the paradoxical effect is hypothesized to be caused by deeper activation of the suppressed thoughts, which makes them more accessible (Harsányi et al., 2014; Stewart et al., 2015). In addition, failed attempts at thought suppression may lead to an increase in negative cognitions and emotions (Bjarnason et al., 2014). Based on these paradoxical effects, thought suppression may be particularly problematic for individuals who attempt to suppress intrusive and unwanted thoughts (Ashton & Boschen, 2011; Koster et al., 2008). Relevant to this point, it appears that thought suppression may be a factor in the development or maintenance of obsessive-compulsive disorder and other mental health disorders including depression, post-traumatic stress and generalized anxiety (Purdon, 1999).

The White Bear Suppression Inventory (WBSI) is perhaps the most widely used instrument to assess the tendency to engage in thought suppression (Wegner & Zanakos, 1994). More specifically, the WBSI is a self-report measure that evaluates a general tendency to suppress unwanted negative thoughts and an inability to ignore unwanted thoughts (Belloch et al., 2009; Jiménez et al., 2015; Schmidt et al., 2009; Wegner & Zankos, 1994). Furthermore, this measure has been found to be positively and significantly associated with symptoms of OCD, depression, and anxiety (e.g., Wegner & Zankos, 1994).

In addition to the original English version of the WBSI, also validated in the United States, the measure has been translated to numerous languages (Schmidt et al., 2009). More specifically, the WBSI has been validated in at least seven additional languages, such as Portuguese, Dutch, English, German, French, Spanish and Turkish. And consistent cross-cultural support has been found for the association between the WBSI and symptoms of OCD, depression and anxiety (Altin & Gençöz, 2009; Jiménez et al., 2015; Muris et al., 1996; Vincken et al., 2012). Despite the fact that the WBSI has been found to be associated with measures of psychopathology, discrepancies in the factor structure have been noted (Schmidt et al., 2009).

Whereas the results of some studies have yielded a single-factor solution such as the original version (Wegner & Zankos, 1994), a number of other studies have arrived at a two-factor solution. Typically, these studies can find a factor that measures the tendency to experience intrusions and a factor that measures the tendency to engage in thought suppression.

Two Spanish versions of the WBSI have been developed and validated. In Spain, the first published study was conducted by Rodríguez et al. (2008) which used the translated version based on the dialect spoken in Spain (European Spanish). The authors explored the structural validity and reliability of the European Spanish version of the WBSI in two community samples from Spain. Reliability and validity were found to be good, and the authors obtained a two-factor solution. These factors were consistent with previous research in which factors that were consistent with intrusion and suppression were obtained.

A Cuban version of the WBSI was developed to reflect the Spanish dialect spoken in Cuba (Cuban Spanish). In the first study, Rodríguez-Martín (2010) examined the latent factor structure and reliability of the Cuban version of the WBSI in a sample of older Cubans. The author found support for a two-factor solution and acceptable reliability. In a follow-up study, Rodríguez-Martín et al. (2014) examined the validity and psychometric properties of the Cuban version of the WBSI community sample. Consistent with previous research, the validity of this version of the WBSI was supported, as the measure was found to be significantly associated with symptoms of anxiety and depression. However, in contrast to Rodríguez-Martín (2010), the authors arrived at a single-factor solution.

Although the WBSI has been translated and validated for the European Spanish and Cuban populations, differences in Spanish dialects across countries may warrant the need for a translated version of the WBSI specifically for the Mexican population (Pallanti, 2008; Treviño-de la Garza et al., 2019). Numerous cultural and linguistic differences between Mexico, Spain and Cuba can be observed, and these differences may increase the probability of misinterpreting specific words or phrases. As it is well known, within a culture, a system of meaning is shared among those who speak the same language or dialect in the same geographic location by using specific words, phrases and sentence structures that become apparent in each culture (Morling & Lamoreaux, 2008; Trandis, 2000). Therefore, it is important to take culture into account and validate the measures with participants within the target context or country (Ruvalcaba et al., 2014).

To illustrate this point, we can compare item three of the WBSI European Spanish version that reads Tengo pensamientos que no puedo evitar with the same item in the Cuban version that reads Tengo pensamientos que no puedo parar. Variations are also likely in the translation of the WBSI to Mexican Spanish. Consequently, the need for a specific version of the WBSI for the Mexican population is very important in order to accurately capture the cultural and linguistic nuances of the Spanish language, as it is spoken in Mexico. This could be accomplished with careful back-to-back translation and a careful selection of specific wording and adaptation of phrases.

In response to this need, the purpose of the current study was to develop the WBSI-Mexican Version (WBSI-MV), and to investigate the validity of this measure in a Mexican sample. Based on other cultural validations of this measure (e.g., Rodríguez et al., 2008; Rodríguez-Martín, 2010), it was hypothesized that the WBSI-MV would report a two-factor structure labeled as intrusion and suppression with good psychometric properties.

Method

Participants

Three hundred and sixty-four students enrolled at university in northern Mexico volunteered to participate in this online survey-based study. Inclusion criteria consisted of participants being 18 years old or older and enrolled in an undergraduate program. Regarding demographic characteristics, the mean age was 20.59 (SD = 1.92), and the sample was 79.4% females and 20% males.

Design and Procedures

The study was internet-based, in which participants completed an online survey packet of self-report measures. Before participating in the study, potential participants were required to complete and sign an informed consent form, as part of the informed consent potential participants were reminded that the involvement was voluntary. Participants who agreed to participate were then asked to complete a demographic questionnaire and the Spanish versions of the measures described below. The study was approved by the university's Institutional Review Board.

Measures

White Bear Suppression Inventory – Mexican Version (WBSI-MV). The WBSI is among the most commonly utilized measures to assess thought suppression (Wegner & Zanakos, 1994). This self-report measure consists of 15 items with response options on a 5-point Likert scale. The WBSI exhibits good psychometric properties, including excellent internal consistency (α = .93), and validity, as the WBSI has been found to be positively and significantly associated with measures of anxiety and depression (Wegner & Zanakos, 1994). For the purpose of the current study, the WBSI was translated from English to Spanish and back-translated from Spanish to English to ensure an accurate translation. Within the translation process, language was adapted and cultural aspects were revised by two professors from the School of Psychology at Mexican university, in order to make the items more accurate and representative to the Mexican population. The back-translated version of the measure was evaluated by native English speakers (please see final version in Appendix A). Cronbach's alphas with the current sample are provided in the results section.

Yale-Brown Obsessive Compulsive Scale Self Report Version (Y-BOCS-SR). The Symptom Severity Scale of the Y-BOCS-SR was administered in this study (Baer, 1991). The Y-BOCS-SR assesses components of symptom severity including the amount of distress, interference, time spent on obsessions or compulsions, and perceived control of obsessions and compulsions. The Y-BOCS-SR severity scale includes a total of seven items to measure obsessions and seven items to evaluate compulsions. Each question has a Likert scale ranging from 0 to 4 (Ólafsson et al., 2010). For the purpose of this study, this measure was also translated from English to Spanish and back-translated. Within the translation process, language was adapted and cultural aspects were revised by two professors from the School of Psychology at Mexican university, in order to make the items more accurate and representative. The back-translated version of the measure was evaluated by native English speakers. The internal consistency of the Y-BOCS-SR in Spanish for the current sample was excellent, α = .93. In the current study this measure was used to assess concurrent validity, as WBSI has consistently been associated with OCD symptoms. The Y-BOCS-SR was used to explore convergent validity, as OCD symptoms have been associated with thought suppression, therefore a degree of convergence was expected.

Thought-Action Fusion Scale (TAFS). The TAFS is a 19-item self-report measure designed to evaluate the construct of thought-action fusion (Shafran et al., 1996). Specifically, thought-action refers to the belief that disturbing and unacceptable thoughts are equal to committing unacceptable behaviors leading to unacceptable actions. The measure has yielded good psychometric properties, including a good internal consistency and a good criterion validity, as the TAFS has been found to be associated with symptoms of OCD (Cougle et al., 2013; Shafran & Rachman, 2004).

There is a Spanish version of this measure published by Jáuregui-Lobera et al., (2013); however, it was written in European Spanish, as opposed to Mexican Spanish. Therefore, this measure was back-to-back translated from English to Spanish. Within the translation process, language was adapted, and cultural aspects were revised by two professors from the School of Psychology at Mexican university, in order to make the items more accurate and representative to the Mexican population. The back-translated version of the measure was evaluated by native English speakers. The internal consistency reported was α =

.94. The TAFS was used to explore convergent validity, as though suppression and thought-action fusion are similar cognitive variables associated with OCD, and a degree of convergence was expected.

Data Analysis Plan

Confirmatory factor analyses (CFAs) were conducted using Mplus version 8.1 to determine the degree of adjustment of the data obtained in the current sample fits with the previously established models. In particular, the purpose of the first CFA was to examine the degree to which the single-factor model obtained by Rodríguez-Martín et al. (2014) and Wegner & Zankos (1994) fits with the current data. The purpose of the second CFA was to examine the degree to which the two-factor model obtained by Rodríguez et al. (2008) fits with the current data. For both CFAs, a WLSMV estimation method was used. Further, pending poor fit with previously established models, an exploratory factor analysis (EFA), was to be conducted with a principal axis extraction method and a Promax rotation. Items with loadings of .30 or greater on a single factor were to be retained.

Following establishment of the factor structure, an assessment of reliability was planned by examining Cronbach's alphas, and an examination of the validity of the WBSI-MV was planned by assessing the magnitude of the association between the WBSI-MV and the designated validation measures (i.e., Y-BOCS-SR and TAFS). The SPSS version 26.0 had been used for the EFA and the bivariate correlations. Finally, qualitative analysis was conducted to determine if the translation process produced substantive differences and improved utility, in order to be used in a Mexican sample, in connection with the previously translated Spanish version of the WBSI.

Results

The single-factor model was found to be a poor fit with the data, χ^2/df = 9.28, CFI = .926, TLI = .914, RMSEA = .14. Further, the two-factor model was also found to be a poor fit with the data, χ^2/df = 7.36, CFI = .945, TLI = .92, RMSEA = .12. Next, based on an Exploratory Factor Analysis, support was found for a two-factor model (see Table 1). Congruent with previous research, one factor was consistent with the theme of thought suppression (labeled "Suppression") and the second was consistent with thought intrusions (labeled "Intrusion"). The Suppression subscale consisted of seven items, and the Intrusion subscale consisted of five items. Two items were dropped due to significant loadings on both factors (See Table 1).

To assess criterion validity, bivariate correlations were conducted to examine association between the subscales of the WBSI-MV and the Y-BOCS-SR. As anticipated, the Suppression subscale, r (356) = .48, p < .001, and Intrusion subscale, r (361) = .54, p < .001, were both found to be significantly associated with the Y-BOCS-SR. Convergent validity was assessed by examining the association between the WBSI-MV and the TAFS. As anticipated, the Suppression subscale, r (356) = .33, p < .001, and Intrusion subscale, r (361) = .27, p < .001, were both found to be significantly and positively associated with the total score of the TAFS.

Finally, a qualitative analysis was conducted in which the items from the European Spanish version of the WBSI were directly compared to items on the WBSI-MV. Differences between the two versions were observed. For example, the

Table 1
Results From a Factor Analysis of the White Bear Suppression Inventory – Mexican Version (WBSI-MV).

Item	Factor 1 (Suppression)	Factor 2 (Intrusion)
1	.56	
8	.58	
10	.77	
11	.91	
12	.55	
13	.79	
14	.73	
15	.50	
2		.71
3		.88
4		.74
5		.77
7		.44

Note. Values based on an exploratory factor analysis with a principal axis factor extraction method and a Promax factor rotation. Factors loadings .30 or above on a single factor were retained for subsequent analyses. Based on this criteria, items 6 and 9 were dropped, as these items yielded loadings of .3 or greater on both factors.

original wording of item 10 reads as Sometimes I stay busy just to keep thoughts from intruding my mind. The European Spanish version employs the word "head" (cabeza), instead of the direct translation of the word "mind" (mente). In contrast, the WBSI-MV uses "mente". Another example relates to item 8, which reads I always try to put problems out of mind, the European Spanish uses the word "quitarme" in relation to putting problems out of mind, while the Mexican Spanish uses the word "apartar" which reflects, in a more accurate way, putting problems out of mind rather than eliminating the problem.

In general, WBSI-MV exhibits differences relative to the European Spanish version of the WBSI, and the translation appears to be successful in taking into account the nuances of the Spanish language, as spoken in Mexico.

Discussion

The purpose of the current study was to examine the psychometric properties of the WBSI translated to Mexican Spanish. Consistent with numerous previous studies, support was found for a two-factor solution, with factors labeled Intrusion and Suppression. Further, also consistent with previous research, the WBSI-MV showed good internal consistency, and support was found for the validity of the WBSI-MV, as both factors were found to be positively and significantly associated with OCD symptoms and thought-action fusion.

The variation in factor structure of the WBSI across studies is noteworthy (Schmidt et al., 2009). The two-factor solution obtained in the current study is generally consistent with the results of the European Spanish version of the WBSI by Rodríguez et al. (2008) and seems to be with other factor analytic studies of the WBSI (see Schmidt et al., 2009). However, a number of studies have obtained a single-factor solution (e.g., Altin & Gençöz, 2009; Jiménez et al., 2015; Vincken et al., 2012). Consequently, more research is needed to better understand the factor structure of the WBSI. This is especially the case for Spanish-speaking individuals, as there is limited research on thought suppression in this population.

To assure accuracy of the translation, the authors of the current study utilized back-to-back translation, and wording and phrases were carefully selected and adapted to reflect a more common linguistic understanding in Mexico. This approach appears to have led to qualitative differences between these two versions, which reiterates the need and utility of a version of the WBSI specifically developed for the Mexican population. For example, one noteworthy difference relates to item 9 with the word "surgiendo" relative to the phrase vienen una y otra vez. In particular, "surgiendo" is a word that more accurately reflects the meaning of the original statement in the Mexican population.

Another difference is observable with item 3, which reads Tengo pensamientos que no puedo detener, in which in the European Spanish version "evitar" refers to a thought that someone may avoid but it eventually returns, as opposed to the translation in the Mexican version in which "detener" is a direct order to stop the unwanted thought. The differences between the two versions reinforce the importance of having a culturally adapted version of the WBSI (Vincken et al., 2012).

Overall, having the WBSI-MV is useful for researchers and clinicians interested in assessing thought suppression in clinical and community samples in Mexico. For example, the WBSI-MV may assist in the identification of OCD and in treatment planning. Moreover, having a WBSI validated in Mexico may serve as an impetus for more research in this country in the areas of anxiety and OCD symptoms.

Some limitations of the present study should be considered and addressed in future research. One of the limitations is the fact that the sample of this study was of undergraduate college students. Consequently, the findings of the current study may not be generalized to non-college students in the Mexican population. As a result, it will be important for subsequent studies to explore the psychometric properties of the WBSI-MV in individuals of various ages, education and SES levels, and from rural and urban areas. In addition, it will expand our current knowledge to have studies that explore the relationship between thought suppression, measured by the WBSI and variables that are closely related to emotional suppression such as coping skills, through measures such as the Coping Strategies Inventory (Schetsche et al., 2022). Further, due to the fact that data collection was conducted with a non-referred sample, it is crucial to do research with clinical samples to better understand thought suppression in patients diagnosed with OCD or anxiety disorders.

In summary, the current findings indicate that the WBSI-MV is a useful and valid tool to assess thought suppression in the Spanish-speaking population in Mexico. The WBSI-MV may also have the potential to provide insight into the underlying mechanisms that lead to the development of anxiety and OCD in the Mexican population. Although the findings are promising, further research should be conducted to continue disseminating and implementing measures, to acquire an accurate depiction of symptoms in culturally diverse communities in order to provide culturally appropriate and effective interventions.

Conflicts of Interest: None.

Funding Statement: This work has been supported by the Research Division of Universidad de Monterrey in Mexico.

References

Altin, M., & Gençöz, T. (2009). Psychopathological correlates and psychometric properties of the White Bear Suppression Inventory in a Turkish Sample. *European Journal of Psychological Assessment, 25*(1), 23-29. https://doi.org/10.1027/1015-5759.25.1.23

Ashton, A. A., & Boschen, M. J. (2011). Thought suppression of multiple personally relevant target thoughts. *Asia Pacific Journal of Counselling and Psychotherapy, 2*(2), 138-150. https://doi.org/10.1080/21507686.2010.544660

Baer, L. (1991). *Getting control: Overcoming obsessions and compulsions*. Boston, MA: Little Brown.

Belloch, A., Morillo, C., & Garcia-Soriano, G. (2009). Strategies to control unwanted intrusive thoughts: Which are relevant and specific in obsessive-compulsive disorder? *Cognitive Therapy and Research, 33*(1), 75-89. https://doi.org/10.1007/s10608-007-9141-2

Bjarnason, R., Emmelkamp, P., Kristjánsson, A., & Ólason, D. (2014). Replacing intrusive thoughts: Investigating though control in relation to OCD symptoms. *Journal of Behavior Therapy and Experimental Psychiatry, 45*, 506-515. https://doi.org/10.1016/j.jbtep.2014.07.007

Cougle, J. R., Purdon, C., Fitch, K. E., & Hawkings, K. A. (2013). Clarifying relations between thought- action fusion, religiosity, and obsessive-compulsive symptoms through consideration of intent. *Cognitive Therapy and Research, 37*(2), 221-231. https://doi.org/10.1007/s10608-012-9461-8

Harsányi, A., Csigó, K., Rajkai, C., Demeter, G., Németh, A., & Racsmány, M. (2014). Two types of impairments in OCD: Obsessions, as problems of thought suppression; compulsions, as behavioral-executive impairment. *Psychiatry Research, 215*(3), 651-658. https://doi.org/10.1016/j.psychres.2013.11.014

Jáuregui-Lobera, I., Santed-Germán, M., Bolaños-Ríos, P., & Garrido-Casals, O. (2013). Spanish version of the Thought-Action Fusion Questionnaire and its application in eating disorders. *Psychology Research and Behavior Management, 6*, 75-86. https://doi.org/10.2147/PRBM.S51183

Jiménez, A., Orgambidez, A., & Pascual, L. (2015). Inventario de Supresión del Pensamiento del Oso Blanco (WBSI): Propiedades psicométricas de la versión portuguesa. *Revista de Psicopatología y Psicología Clínica, 20*, 125-134. https://doi.org/10.5944/rppc.vol.20.num.2.2015.15167

Koster, E. H. W., Soetens, B., Braet, C., & De Raedt, R. (2008). How to control a white bear? Individual differences involved in self-perceived and actual thought-suppression ability. *Cognition and Emotion, 22*(6), 1068-1080. https://doi.org/10.1080/02699930701616591

Morling, B., & Lamoreaux, M. (2008). Measuring culture outside the head: A meta-analysis of individualism-collectivism in cultural products. *Personality and Social Psychology Review, 12*(3), 199-221. https://doi.org/10.1177/1088868308318260

Muris, P., Merckelbach, H., & Horselenberg, R. (1996). Individual differences in thought suppression The White Bear Suppression Inventory: Factor structure, reliability, validity and correlates. *Behaviour Research and Therapy, 34*(5-6), 501-513. https://doi.org/10.1016/0005-7967(96)00005-8

Ólafsson, R. P., Snorrason, Í., & Smári, J. (2010). Yale-Brown Obsessive-Compulsive Scale: Psychometric properties of the self-report version in a student sample. *Journal of Psychopathology and Behavioral Assessment, 32*, 226-235. https://doi.org/10.1007/s10862-009-9146-0

Pallanti, S. (2008). Transcultural observations of obsessive-compulsive disorder. *The American Journal of Psychiatry, 165*(2), 169-170. https://doi.org/10.1176/appi.ajp.2007.07111815

Pica, G., Amato, C., Mauro, R., & Pierro, A. (2015). Individual differences in preference for thought suppression: Components and correlates of the White Bear Suppression Inventory. *TPM-Testing, Psychometrics, Methodology in Applied Psychology, 22*(1), 31-42.

Purdon, C. (1999). Thought suppression and psychopathology. *Behaviour Research and Therapy, 37*(11), 1029-1054.

https://doi.org/10.1016/S0005-7967(98)00200-9

Rodríguez, M. G., Avero-Delgado, P., Rovella, A. T., & Cubas-León, R. (2008). Structural validity and reliability of the Spanish version of the White Bear Suppression Inventory (WBSI) in a sample of the general Spanish population. *The Spanish Journal of Psychology, 11*(2), 650-659. https://doi.org/10.1017/s1138741600004650

Rodríguez-Martín, B. C. (2010). Estructura factorial y confiabilidad del White Bear Suppression Inventory en una muestra de adultos mayores de las provincias centrales de Cuba. *Revista Cubana de Psicología, 23*(1), 40-45. Recuperado de https://biblat.unam.mx/es/revista/revista-cubana-de-psicologia

Rodríguez-Martín, B. C., Molerio-Pérez, O., Martínez-Rodríguez, L., González-Paneca C. L., & Navarro-Otero, S. M. (2014). Estructura factorial, confiabilidad y validez del Inventario de Supresión del Oso Blanco en adultos cubanos. *Alternativas Cubanas en Psicología, 2*(6), 92-101.

Ruvalcaba-Romero, N. A., Gallegos-Guajardo, J., Lorenzo-Alegría, M., & Borges del Rosal, Á. (2014). Propiedades psicométricas del Inventario de Competencias Socioemocionales para adolescentes (EQi- YV) en población mexicana. *Revista Evaluar, 14*, 1-14. https://doi.org/10.35670/1667-4545.v14.n1.8409

Schetsche, C., Jaume, L.C., & Azzollini, S. (2022). Desarrollo de una versión breve del Coping Strategies Inventory. *Revista Evaluar, 22*(1) 1-16. https://doi.org/10.35670/1667-4545.v22.n1.37412

Schmidt, R. E., Gay, P., Courvoisier, D., Jermann, F., Ceschi, G., David, M., Brinkmann, K., & Van der Linden, M. (2009). Anatomy of the White Bear Suppression Inventory (WBSI): A review of previous findings and a new approach. *Journal of Personality Assessment, 91*(4), 323-330. https://doi.org/10.1080/00223890902935738

Shafran, R., & Rachman, S. (2004). Thought-action fusion: A review. *Journal of Behavior Therapy and Experimental Psychiatry, 35*(2), 87-107. https://doi.org/10.1016/j.jbtep.2004.04.002

Shafran, R., Thordarson, D. S., & Rachmann, S. (1996). Thought-action fusion in obsessive-compulsive disorder. *Journal of Anxiety Disorders, 10*, 379-391. https://doi.org/10.1016/0887-6185(96)00018-7

Stewart, I., Hooper, N., Walsh, P., O'Keefe, R., Joyce, R., & McHugh, L. (2015). Transformation of thought suppression functions via same and opposite relations. *The Psychology Record, 65*, 375-399. https://doi.org/10.1007/s40732-014-0113-0

Treviño-de la Garza, B., Berman, N., Fisak, B., Ruvalcaba-Romero, N., & Gallegos-Guajardo, J. (2019). Validation of The Dimensional Obsessive-Compulsive Scale for Mexican population. *Journal of Obsessive-Compulsive and Related Disorders, 21*, 13-17. https://doi.org/10.1016/j.jocrd.2018.11.006

Triandis, H. C. (2000). Culture and conflict. *International Journal of Psychology, 35*(2), 145-152. https://doi.org/10.1080/002075900399448

Vincken, M. J. B., Meesters, C., Engelhard, I. M., & Schouten, E. (2012). Psychometric qualities of the White Bear Suppression Inventory in a Dutch sample of children and adolescents. *Personality and Individual Differences, 52*(3), 301-305. https://doi.org/10.1016/j.paid.2011.10.023

Wegner, D. M., & Zanakos, S. (1994). Chronic thought suppression. *Journal of Personality, 62*(4), 615-640. https://doi.org/10.1111/j.1467-6494.1994.tb00311.x

Appendix A

White Bear Suppression Inventory for Mexican Version (WBSI-MV)
Inventario de Supresión del Pensamiento del Oso Blanco Versión Mexicana

Por favor indica el grado concordancia o discrepancia con los elementos de la siguiente escala:

En completo desacuerdo	En desacuerdo	Neutral	De acuerdo	De acuerdo completamente
1	2	3	4	5

1. Hay cosas en las que prefiero no pensar.
2. Algunas veces me pregunto por qué tengo los pensamientos que tengo.
3. Tengo pensamientos que no puedo detener.
4. Hay imágenes que vienen a mi mente que no puedo borrar.
5. Mis pensamientos vuelven a una misma idea frecuentemente.
6. Desearía poder dejar de pensar en ciertas cosas.
7. Mi mente se acelera tanto algunas veces que desearía poder detenerla.
8. Siempre trato de apartar los problemas de mi mente.
9. Existen pensamientos que continúan surgiendo en mi cabeza.
10. Algunas veces me mantengo ocupado para evitar que ciertos pensamientos interfieran en mi mente.
11. Hay cosas en las que prefiero no pensar.
12. Algunas veces solo quisiera dejar de pensar.
13. Hago cosas con frecuencia para distraerme de mis propios pensamientos.
14. Tengo pensamientos que trato de evitar.
15. Hay muchos pensamientos que tengo que no comparto con nadie.

Scoring notes
(1) Suppression subscale (sum of items 1,8,10,11,12,13,14,15).
(2) Intrusion subscale (sum of items 2,3,4,5,6,7).
(3) Items 6 and 9 were dropped due to dual loadings (significant loadings on both factors).

2023, Vol. 23, No. 2
ISSN 1667-4545
Recuperado de https://revistas.unc.edu.ar/index.php/revaluar

Laboratorio de Evaluación Psicológica y Educativa
Facultad de Psicología - Universidad Nacional de Córdoba

Teacher's Achievement Verbalizations Questionnaire in Oral Exams

Cuestionario de Verbalizaciones de Logro Docentes en Exámenes Orales

Javier Sánchez-Rosas *[1] , Sergio Alexis Dominguez-Lara[2] , Luis Alberto Furlan[3] ,
Mirai Okinishi[3]

1 - Catholic University of Temuco, Temuco, Chile.
2 - Private University Norbert Wiener, Lima, Perú.
3 - National University of Cordoba, Córdoba, Argentina.

Recibido: 20/02/2023 **Revisado:** 24/04/2023 **Aceptado:** 18/05/2023

Abstract

At the moment, there is no measurement instrument to assess verbal expressions and phrases of feedback about students' achievement issued by teachers in the presence of one or more students during an oral exam. This article reports the design, structural validity, reliability and external validity of the Teacher Achievement Verbalizations in Oral Exams Questionnaire (TAVQ), which assesses several teachers' verbalizations from the perspective of students in oral exams. The structural validity, reliability, and external validity were evaluated in a sample of university students (N = 252) from Argentina. Several plausible measurement models were specified based on the dimensions of valence, object focus, and temporal frame, which were tested through confirmatory factor analysis and bifactor analysis. Two scales that measure with very good reliability positive and negative verbalizations related to achievement expressed by teachers during oral exams were validated. These verbalizations showed appropriate relationships with achievement emotions and academic performance. The need for future studies and practical implications are discussed.

Keywords: *test anxiety, teacher behavior, feedback, oral exams, achievement emotion*

Resumen

Hasta el momento, no existe un instrumento de medición que evalúe las expresiones verbales y comentarios sobre el logro de los alumnos que emiten los docentes en presencia de uno o varios estudiantes durante un examen oral. Este artículo informa la construcción, validez estructural, confiabilidad y validez externa del cuestionario de verbalizaciones de logro docentes en exámenes orales (TAVQ). El mismo evalúa varias verbalizaciones del docente desde la perspectiva de los estudiantes en exámenes orales. La validez estructural, confiabilidad y validez externa fueron evaluadas en una muestra de estudiantes universitarios (N = 252) de Argentina. Se especificaron varios modelos de medición plausibles basados en las dimensiones de valencia, foco y marco temporal, que fueron testeados mediante análisis factorial confirmatorio y bifactor. Se validaron dos escalas que miden con muy buena confiabilidad verbalizaciones de logro positivas y negativas expresadas por docentes durante los exámenes orales. Estas verbalizaciones mostraron relaciones apropiadas con emociones de logro y rendimiento académico. Se discute la necesidad de estudios futuros e implicancias prácticas.

Palabras clave: *ansiedad ante los exámenes, comportamiento docente, comentarios, exámenes orales, emoción de logro*

***Correspondence to:** Javier Sánchez-Rosas, Department of Psychology, Faculty of Health Sciences, Catholic University of Temuco, Manuel Montt 56, Temuco, Chile. E-mail: sanchezrosasjavier@gmail.com

How to cite: Sánchez-Rosas, J., Dominguez-Lara, S. A., Furlan, L. A., Okinishi, M. (2023). Teacher's Achievement Verbalizations Questionnaire in Oral Exams (TAVQ). *Revista Evaluar, 23*(3), 67-82. Retrieved from https://revistas.unc.edu.ar/index.php/revaluar

Authors' Notes: Javier Sánchez-Rosas https://orcid.org/0000-0002-7669-8981, Sergio Alexis Dominguez-Lara https://orcid.org/0000-0002-2083-4278, Luis Alberto Furlan https://orcid.org/0000-0002-8415-0596, Mirai Okinishi https://orcid.org/0000-0003-2979-6818.

Participaron en la edición de este artículo: Abigail Pérez, Stefano Macri, Eugenia Barrionuevo, Rita Hoyos, Florencia Ruiz, Ricardo Hernández, Jorge Bruera.

Introduction

Oral exams are frequent in higher level education and are designed to assess students' understanding of a subject, as well as the ability to articulate ideas and knowledge effectively (Hazen, 2020; Theobold, 2021). Whoever faces an oral exam must develop skills to respond adequately to a double challenge in a stressful context that can activate several achievement emotions (Furlan & Sánchez-Rosas, 2018). On the one hand, the evaluated content must be communicated in a clear, fluent and organized way, using the specific terms of the subject in an appropriate manner. On the other hand, the student must sustain an asymmetric interpersonal interaction, while processing information related to his/her performance received through the gestural and verbal language of his/her evaluator (Burić, 2015; Gardner & Giordano, 2023; Guo et al., 2022; Puertas-Molero et al., 2022; Putwain et al., 2017). The messages related to achievement can affect control and value appraisals of the ongoing activity and their outcomes, and promote increased achievement emotions (Goetz et al., 2018).

Achievement emotions are defined as emotions that are directly linked to achievement activities or achievement outcomes (Pekrun, 2018, 2021). Achievement emotions are present daily in the academic environment and have the ability to affect students' thoughts, motivation and actions in evaluative situations (Furlan & Sánchez-Rosas, 2018; Pekrun, 2018; Pekrun et al., 2023; Rojas-Torres et al., 2022; Sánchez-Rosas & Furlan, 2017). These emotions are activated by control-value appraisals and the learning context would contribute to their activation by affecting these appraisals (Pekrun, 2018, 2021).

Beyond the existence of individual causes, there is a growing interest in knowing the environmental characteristics that impact students' emotional experiences (Dewaele et al., 2018, 2019; Lei et al., 2018; Pekrun et al., 2023; Raccanello et al., 2018; Sánchez-Rosas et al., 2019; Ventura-León et al., 2022; Yang et al., 2021). In this regard, teachers, through their behaviors in the classroom, (Sánchez-Rosas et al., 2016; Sánchez-Rosas & Esquivel, 2016), would play a crucial role in oral exam situations by affecting control-value appraisals of exams (Burić, 2015; Putwain et al., 2022; Reilly & Sánchez-Rosas, 2021).

One specific case on teaching behaviors is teachers' achievement verbalizations in oral exams, which can increase the demands of the task, provide clarity and structure to the exam, or transmit messages related to achievement in terms of success or failure. These achievement verbalizations could impact students' emotions and other important aspects of student performance (Pekrun et al., 2023). When teachers use negative, judgmental, or frightening language, it can generate negative emotions in students that affect students' motivation and academic behavior (Apto et al., 2017; Putwain et al., 2017, 2023). In contrast, if teachers use positive and motivating language, it can lead to less negative emotional, motivational, and behavioral consequences.

Although it is possible to envision the effects of these positive or negative achievement verbalizations on achievement emotions, instruments that allow measuring teachers' achievement verbalizations in oral exams are still nonexistent. To address this gap and provide for related research, a study that seeks to develop a tool to assess various teacher achievement verbalizations in oral exams and to analyze some of their psychometric properties is reported. Specifically, the dimensional structure of a set of items is analyzed as well as the reliability of the measurements made by the resulting scales, and the relationship that teachers' achievement verbalizations have with achievement emotions during oral exams is also tested.

When developing the instrument, an empirical-rational strategy was followed, paying attention to the ability of achievement verbalizations to affect control and value appraisal and, consequently, to activate achievement emotions (Pekrun et al., 2023). In addition, various measurement models were analyzed based on possible combinations according to the valence and object focus of the emotion that achievement verbalizations could activate. Also, although a large pool of items was developed, the final instrument is a brief and practical version that will shorten the length of research protocols in studies with many variables and avoid the presence of less representative items of the construct.

Achievement Emotions and Teacher Behaviors

The CVT offers a frame of reference to define teachers' achievement verbalizations that activate achievement emotions in oral exams, while building scales and validating instruments (Pekrun, 2018, 2021; Pekrun et al., 2023). It is important to note that the CVT establishes that control (e.g., self-efficacy) and value (e.g., task value) appraisals are the direct causes of the activation of achievement emotions. This means that emotions are induced when the individual feels in control of, or out of control of, activities and outcomes that are subjectively important. In turn, the emotions activated in achievement situations, such as oral exams, can be distinguished by their valence (positive vs. negative) or by the object focus of the emotion (activity or outcomet); emotions can even be distinguished as current (activity), prospective (future outcomes), or retrospective (past outcomes) emotions.

The control-value theory postulates that the affective impact of social environments is mediated by control and value appraisals. Accordingly, it is assumed that the features of environments that deliver information related to controllability and academic values are of critical importance for students' emotions. Important variables include quality of instruction, induction of values, autonomy support, goal structures and achievement-related expectancies of significant others, as well as feedback and consequences of achievement (Pekrun, 2018).

Understanding the role of immediate environmental factors in achievement situations, such as teacher behavior or teacher feedback (Apto et al., 2017; Awad-Igbaria et al., 2022; Frenzel et al., 2021; Narciss et al., 2022) is important because it allows us to understand how they influence the students' constitution of achievement beliefs and expectations in oral exams. Various teaching behaviors have been considered when analyzing their relationship with control-value appraisals and emotions (Becker et al., 2014; Lazarides & Buchholz, 2019; Lei et al., 2018; Sánchez Rosas et al., 2016; Westphal et al., 2018). Some of these behaviors can be categorized as non-verbal behaviors (Derakhshan et al., 2023; Guo et al., 2022; Juma et al., 2022; Puertas-Molero et al., 2022) and their influence can sometimes be ambiguous depending on students' interpretation: space management, gestures, body language, position and body orientation, gaze, facial expression, and paralinguistic features such as tone of voice and rhythm.

Teacher's achievement verbalizations in oral exams

Teachers' achievement verbalizations are messages that have a much more precise capacity to transmit information related to achievement than non-verbal behaviors and, consequently, their effect on appraisals and emotions is clear-er (Putwain et al., 2017, 2022, 2023). Teachers' achievement verbalizations in oral exams are understood here as verbal expressions, phrases, and feedback on achievement issued by teachers in the presence of one or more students (Narciss et al., 2022; Putwain et al., 2017, 2022, 2023). If, for instance, before starting the exam the teacher warns that the exam will be difficult, this enhances the perceived difficulty of the exam and can elicit anxiety (Putwain et al., 2017). If, by contrast, the teacher gives negative feedback on the current performance (He/she says: *How could you not know that?*), it can activate shame (Apto et al., 2017). By contrast, if the teacher gives positive feedback about current performance or provides support to continue responding, pride and enjoyment can be activated (Pekrun et al., 2023).

As it can be seen, it is possible to think that some verbalizations are more associated with one emotion than others or that they are even related to several emotions simultaneously. For example, arbitrary questions or humiliating expressions about skill level or knowledge can clearly mobilize ideas of unfair treatment and anger, but also anxiety and hopelessness by inducing loss of control. Some verbalizations can even anticipate success before the exam starts, which can increase hope regarding the possibility of obtaining a positive result and activate enjoyment of the situation.

Items development of the (TAVQ)

Item construction was based on the aforementioned theoretical aspects and a preliminary exploratory study (Sánchez-Rosas, 2016). This exploratory study analyzed the occurrence of various teachers' achievement verbalizations in oral exams associated with some achievement emotions.

Seven sets of items were developed, each one initially thought to be closely related to one of seven possible emotions (enjoyment, anger, pride, hope, anxiety, shame and hopelessness). In this way, one set of items would evaluate verbalizations that could activate enjoyment (e.g., *The teacher affirms that the exam is one more instance of learning*), while another item would evaluate hopelessness, (e.g., *The teacher says that even if I try hard, I will not be able to improve my exam performance*). The selection of these items associated with these seven emotions would cover verbalizations that were thought to be associated with frequent emotions in exams and with positive-negative emotions of activity (enjoyment and anger) and of past or future outcomes (hope, anxiety and hopelessness, pride, and shame). It is important to note that these items would not evaluate emotions but verbalizations that would be believed to be discernible from one another and that their grouping could be due to their possible association with a specific emotion. However, it must be recognized that the same verbalization could simultaneously lead to experiencing emotions that are empirically difficult to separate due to their valence, object focus, or time frame (e.g., anxiety, hopelessness; Pekrun et al., 2023; Sánchez-Rosas, 2016). The item design contemplated the specific achievement context (oral exams) and the temporal nature (before/beginning, during, after/end) predominant in the achievement context that triggers each emotion (Pekrun, 2018, 2021). This is because emotions can be activated before, during and after the achievement activity (questions, problem situations) or before and after the achievement outcomes (anticipations, feedback). However, some emotions are more prospective (hopelessness), others more retrospective (pride), and others more concurrent (anger). Therefore, when writing the items, we sought to give relevance to the temporal aspect of the achievement. Thus, for example, an item that was thought to be associated with anxiety included verbalizations by the teacher mentioning the difficulty of the situation at the beginning of the exam or conveying uncertainty about the appropriateness of the responses during the exam. In contrast, to evaluate a verbalization associated with shame, an item was written in which the teacher explained to other people that the student failed to respond.

Method

Participants

Two hundred fifty-two students in different academic programs (72% were studying psychology) from the National University of Córdoba, Argentina, participated. The participants were between the ages of 18 and 55 (M = 27.70, SD = 8.81), in their first to fifth year of studies (5th year = 40%), and 84% of the participants were women.

Instruments

Teacher's achievement verbalizations in oral exams. For the assessment of teacher's achievement verbalizations in oral exams, 56 preliminary items were used, which were distributed according to the measurement model tested. For example, if oblique model of seven dimensions B was tested, items were distributed in groups of eight items depending on whether they had been designed to evaluate their association with enjoyment (*The teacher greets at the beginning of the exam*), anger (*The teacher makes questions about topics that are not on the syllabus*), anxiety (*Before starting, the teacher warns that the exam will be difficult*), shame (*At the end of the exam, the teacher says: I thought you had prepared better*), hope (*Before starting, the teacher assures me that I have the enough ability to pass the exam*), hopelessness (*Before starting, the teacher assures me that the exam will be very difficult*) and pride (*When the exam is over, the teacher congratulates me on my performance*). The final version of the instrument includes eight items of positive verbalizations (*When the exam is over, the teacher compliments the way I prepared myself for the exam*) and eight items of negative verbalizations (*While I take the exam, the teacher reproaches: You should have already known that!*). Each of the items is answered on a 5-point Likert scale (1 = *almost never*, 5 = *almost always*) to describe the frequency with which teachers make various comments or verbalizations related to achievement in oral exams. Although the response scale used evaluated the typical experience in oral exams, it is possible to evaluate the experience in a particular exam with slight modifications in the response instruction.

Achievement emotions. Seven items were used to assess the emotions of enjoyment (*I enjoy taking the exam*), anger (*I get angry*), anxiety (*I am very nervous*), shame (*I feel ashamed*), hope (*I am very confident*), hopelessness (*I feel hopeless*) and pride (*I am very pleased with myself*). Each of the items is answered on a 5-point Likert scale (1 = *almost never*, 5 = *almost always*) to describe the frequency with which students experience these emotions in oral exams (Sánchez-Rosas, 2015).

Academic performance. Students' performance was measured by assessing their average grades attained over the academic career.

Procedures

The set of items was included in an online survey to which questions about gender, age, career, and current academic year were added. The survey included an invitation to participate, the objectives of the study and an informed consent form. It was conducted through a platform that takes online surveys and the invitation was made through social networks.

Data Analysis

Measurement models of teachers' achievement verbalizations in oral exams. Considering the valence, object focus, and temporality of emotions, seven alternative measurement models of teacher achievement verbalizations in oral exams were specified: (a) uni-dimensional model A, in which all items load on a single factor; (b) oblique model of seven dimensions B, in which the seven sets of designed items load their respective latent factors with specific emotions; (c) bi-dimensional model C, in which the items of positive verbalizations and negative verbalizations load their respective latent factor; (d) bi-dimensional model D, in which the items of activity (enjoyment + anger) and outcomes (anxiety + shame + hopelessness + pride) load their respective latent factor; (e) tri-dimensional model E, in which the items of activity (enjoyment + anger), of positive (hopelessness + pride) and negative (anxiety + shame) outcomes load their latent factor respectively; (f) tri-dimensional model F, in which the items of activity (enjoyment + anger), prospective (anxiety + hopelessness + hope) and retrospective (shame + pride) outcomes load their latent factor respectively; (g) bifactor model G, in which all the items load on a general factor and the items with positive and negative verbalizations load on their respective latent factor.

Preliminary analysis. The approximation to univariate normality was analyzed through the magnitude of the skewness and kurtosis of the items, and they were considered acceptable if they were less than 2 and 7, respectively (Finney & DiStefano, 2006). In turn, for multivariate normality, the Mardia multivariate kurtosis coefficient was used (G2 < 70; Pérez et al., 2013).

Assessment of measurement models. The different measurement models were assessed based on the magnitude of the fit indices such as the CFI (Comparative Fit Index > .90; McDonald & Ho, 2002), the RMSEA (Root Mean Square Error of Approximation < .08; Browne & Cudeck, 1993), and the WRMR (Weighted Root Mean Square Residual Index < 1.00; DiStefano et al., 2018). In turn, factorial loads (> .50; Dominguez-Lara, 2018) and interfactorial correlations were also considered in order to detect potential cases of factorial redundancy (ϕ > .80; Brown, 2015).

If the bifactor model obtains a favorable fit, complementary indicators will be assessed to evaluate the representativeness of the general factor (ω_h, ω_{hs}, and ECV; Rodriguez et al., 2016), while values less than .30 for ω_{hs} reaffirm it.

Short version of the TAVQ. Based on the criteria mentioned above, the best measurement model was chosen, with which a brief version was designed, selecting the items whose factor loadings are greater than .60 by progressively eliminating those that were below that requirement. After designing it, the equivalent with the extended version was analyzed using the corrected Pearson correlation coefficient, since both versions share items, and a corrected correlation above .70 was expected to conclude the equivalence of the versions.

Reliability

Finally, the reliability of the brief version was evaluated both at the score level (α > .70; Ponterotto & Charter, 2009) and the construct level (ω > .70; Hunsley & Marsh, 2008).

Estimation and software

A series of confirmatory factor analyses were implemented in order to assess the measurement models proposed in the introductory section. The weighted least squares means and variance adjusted estimation method (WLSMV) and

Table 1
Descriptive statistics of the items.

	M	SD	g_1	g_2		M	SD	g_1	g_2
Item 1	1.93	1.18	0.97	-0.20	Item 29	1.84	1.04	0.89	-0.38
Item 2	2.07	1.23	0.80	-0.59	Item 30	1.28	0.78	3.18	10.21
Item 3	2.31	1.27	0.57	-0.82	Item 31	2.05	1.22	0.80	-0.51
Item 4	3.15	1.27	-0.05	-1.03	Item 32	1.32	0.75	2.53	6.24
Item 5	1.78	1.00	1.18	0.66	Item 33	4.25	1.00	-1.19	0.53
Item 6	2.65	1.20	0.24	-0.75	Item 34	2.65	1.08	0.06	-0.56
Item 7	1.81	1.06	1.11	0.21	Item 35	3.04	1.29	-0.17	-1.06
Item 8	1.53	0.94	1.79	2.56	Item 36	2.86	1.21	0.08	-0.88
Item 9	1.96	1.16	0.92	-0.31	Item 37	2.67	1.08	-0.01	-0.64
Item 10	2.54	1.26	0.23	-0.95	Item 38	2.67	1.05	0.17	-0.45
Item 11	2.05	1.36	0.86	-0.73	Item 39	2.32	1.17	0.29	-0.97
Item 12	2.53	1.24	0.29	-0.88	Item 40	2.52	1.09	0.03	-0.91
Item 13	2.36	1.30	0.50	-0.93	Item 41	2.33	1.20	0.48	-0.68
Item 14	1.86	1.15	1.25	0.62	Item 42	2.00	1.04	0.81	-0.03
Item 15	2.28	1.09	0.43	-.62	Item 43	1.56	0.91	1.61	2.01
Item 16	1.86	1.05	1.07	0.37	Item 44	2.23	1.23	0.47	-1.10
Item 17	2.13	1.12	0.53	-0.81	Item 45	2.31	1.10	0.32	-0.74
Item 18	2.31	1.18	0.37	-0.97	Item 46	1.59	0.89	1.33	0.76
Item 19	1.69	0.99	1.28	0.79	Item 47	1.94	1.06	0.83	-0.27
Item 20	1.44	0.84	1.99	3.38	Item 48	1.90	1.03	0.89	-0.08
Item 21	1.81	1.08	1.22	0.60	Item 49	1.82	1.12	1.23	0.58
Item 22	1.91	1.17	1.01	-0.16	Item 50	1.45	0.87	2.37	5.80
Item 23	2.44	1.25	0.41	-0.88	Item 51	1.39	0.84	2.40	5.81
Item 24	2.23	1.20	0.60	-0.67	Item 52	1.71	1.01	1.33	0.98
Item 25	1.69	1.02	1.33	0.82	Item 53	2.20	1.10	0.46	-0.72
Item 26	1.84	1.13	1.17	0.41	Item 54	1.80	1.01	1.08	0.46
Item 27	1.87	1.14	1.08	-0.02	Item 55	1.72	1.04	1.28	0.62
Item 28	1.91	1.16	0.93	-0.31	Item 56	2.42	1.08	0.28	-0.59

Note. M= Mean; SD= Standard Deviation; g_1= Asymmetry; g_2= Kurtosis; items 1 to 8= anxiety; items 9 to 16= shame; items 17 to 24= anger; items 25 to 32= hopelessness; items 33 to 40= enjoyment; items 41 to 48= pride; items 49 to 56= hope. The numbering corresponds to the database ordered for analytical purposes. The items were presented to the participants in order.

the polychoric correlation matrix between items were used. For this purpose, the software Mplus v. 7 (Muthén & Muthén, 2012).

Results

Evidence of validity related to internal structure: Preliminary analysis

Most of the items present magnitudes of skewness and kurtosis that are close to univariate normality (e.g., item 1), but in other cases they far exceed the established limits (e.g., item 30) (Table 1). Likewise, G2 is above what is suggested to conclude on multivariate normality (G2 = 225.530).

Evaluation of measurement models

Regarding the analysis of the internal structure, four models (A, D, E and F) presented fit indices with unacceptable magnitudes. Likewise, the bifactor model (model G) did not achieve a better fit than the other models, so the analysis of the complementary indicators was not continued.

Those that had a more acceptable adjustment were the model of seven oblique factors (model B) and two factors (model C), although in the case of the first, the interfactorial correlation was high among those factors that corresponded to the same type of verbalization, either negative (ϕ_{mean} = .949) or positive (ϕ_{mean} = .830). In this sense, the two-factor model was considered the most parsimonious one and the one which best represents the evaluated construct.

Short version of TAVQ

The short version was prepared based on the two-factor model (negative verbalizations and positive verbalizations) by gradually discarding the items whose factorial loads were less than .60 in their respective factors.

Table 2
TAVQ measurement models.

	CFI	RMSEA	IC_{RMSEA} 90%	WRMR
Model A	.661	.103	.100, .106	2.468
Model B	.835	.072	.069, .076	1.785
Model C	.831	.073	.070, .076	1.846
Model D	.696	.108	.105, .111	2.336
Model E	.753	.097	.094, .101	2.099
Model F	.665	.103	.100, .106	2.455
Model G	.749	.092	.089, .095	2.062

Note. Model A= uni-dimensional model; Model B= oblique model of seven dimensions; Model C= bi-dimensional model of positive and negative factors; Model D= bi-dimensional model of activity and outcomes factors; Model E= tri-dimensional model of activity and positive and negative outcomes factors; Model F= tri-dimensional model of activity and prospective and retrospective factors; Model G= bifactor model.

Table 3
TAVQ Factorial Loadings and Correlations.

Item		Factorial loadings
	Negative verbalizations	
13	While I am taking the exam, the teacher reproaches: You should have already known that! Mientras rindo, reprocha ¡Eso ya lo deberías saber!	.78
19	The teacher makes negative comments about my skill. Hace comentarios negativos sobre mi capacidad.	.78
20	The teacher makes fun of what I say. Se burla de lo que digo.	.77
27	While I am taking the exam, the teacher says: So far, you could not answer anything right. Mientras rindo dice: hasta el momento no pudiste responder nada bien.	.86
28	While I am taking the exam, the teacher assures me that my level of knowledge is not enough to pass. Mientras rindo asegura que mi nivel de conocimiento es insuficiente para aprobar.	.84
29	The teacher assures me that, with the knowledge I have, I will not be able to pass the exam. Asegura que con lo que sé no podré aprobar el examen.	.84
31	The teacher maintains that it makes no sense to keep asking me. Sostiene que no tiene sentido seguir preguntándome.	.78
32	The teacher says that even if I make an effort, I won't be able to improve my performance. Dice que aunque me esfuerce, no podré mejorar mi desempeño en el examen.	.73
	Positive verbalizations	
41	When the exam is over, the teacher congratulates me on my performance. Al terminar el examen, me felicita por mi desempeño.	.82
42	When the exam is over, the teacher compliments the way I prepared myself for the exam. Al terminar el examen, elogia la forma en que me preparé.	.79
44	When the exam is over, the teacher encourages me to keep on the same track. Al terminar el examen, me alienta a seguir así.	.83
47	The teacher says that my answer is excellent. Dice que mi respuesta es excelente.	.71
51	Before we start the exam, the teacher asserts that he/she has confidence in me. Antes de comenzar, dice que confía en mí.	.60
52	Before we start the exam, the teacher says I will do great just like everyone else. Antes de comenzar, asegura que me va a ir bien como a todos.	.64
54	The teacher asserts that, by what I seem to know, for sure I will do well. Afirma que, por el nivel de conocimiento que demuestro, seguro me va a ir bien.	.64
56	The teacher thinks that I have a good understanding of the topics. Considera que tengo un buen dominio de los temas.	.70
	Factor Correlations	-.36

Regarding the equivalence between the long and short version of the negative verbalizations factor, the initial correlation was .91, and a post correction of .88 was obtained; and in the case of positive verbalizations, initially it reached .92 initially, and when implementing the correction, it decreased to .86. In both cases, it is concluded that the long and short versions are equivalent.

Reliability

Finally, in relation to reliability, adequate values were obtained both at the construct level and for scores of the negative verbalizations dimension (ω = .936; α = .895) and positive verbalizations (ω = .898; α = .856).

Evidence of validity by association with other variables

The association of verbalizations with achievement emotions and academic performance was significant in almost all cases (Table 4), with moderate and low results.

Discussion

This study presented the development of an instrument to evaluate teachers' achievement verbalizations in oral exams and its psychometric properties assessed through internal structure, reliability and criterion related variables. Altogether, the results provide psychometric evidence for an instrument that allows to evaluate teachers' achievement of positive and negative verbalizations in oral exams. These verbalizations include, among others, messages that anticipate difficulty or uncertainty in the exam, indicate poor performance, imply an arbitrary attitude or promote control of the exam. In addition, these verbalizations are related to the activation of various emotions and academic performance.

The progress presented here constitutes the focus on an important area of research on academic emotions, a topic that is not much studied at the local level (specifically, in the research on teachers' verbalizations that affect the activation of student's emotions during oral exams) (Narciss et al., 2022; Putwain et al., 2017, 2022, 2023).

Table 4
Association of teachers' verbalizations with achievement emotions and academic performance.

	1	2	3	4	5	6	7	8	9	10
1. NV	-									
2. PV	-.27***	-								
3. enjoyment	-.28***	.43***	-							
4. pride	-.30***	.39***	.56***	-						
5. anger	.27***	-.08	-.08	-.18**	-					
6. anxiety	.07	-.12*	-.41***	-.24***	.15*	-				
7. shame	.20**	-.24***	-.41***	-.42***	.19**	.43***	-			
8. hope	-.15*	.26***	.46***	.58***	-.09	-.25***	-.42***	-		
9. hopelessness	.33***	-.27***	-.39***	-.55***	.25***	.23***	.51***	-.41***	-	
10. GPA	-.29***	.29***	.16**	.35***	-.14*	-.04	-.25***	.16**	-.32***	-

Note. * $p < .05$, ** $p < .01$, *** $p < .001$. GPA= Grade Point Average; NV= negative verbalizations; PV= positive verbalizations.

Item development and evidence of validity and reliability

The item design intended to incorporate the dimensions of the valence, object focus and temporal frame (Pekrun, 2018, 2021) of the emotions to which teachers' achievement verbalizations would be associated. In turn, the items were designed considering aspects that would be core to each achievement emotion, for example, uncertainty about the domain of the exam (Putwain et al., 2017), exposure of the error (Apto et al., 2017) or certainty about a positive result (Pekrun et al., 2023).

Considering the dimensions and these central aspects, plausible models for measuring teachers' achievement verbalizations in oral exams were specified. Seven measurement models were evaluated and the most favorable adjustments were obtained for a model with seven oblique but highly related dimensions and a more parsimonious model with two dimensions that grouped positive and negative verbalizations. As anticipated, the same verbalization could simultaneously activate various emotions that are empirically difficult to separate due to their valence, emotional focus, or time frame (e.g., anxiety and hopelessness; Pekrun et al., 2023; Sánchez-Rosas, 2016). Some of the evaluated models considered the time frame of the verbalizations; however, they did not show a good fit to the data. This difficulty in discriminating the time frame may be due to the fact that the items refer to verbalizations that are performed at the beginning and at the end of the exam, instead of being performed before or after it. In this way, all the items would have been constructed as concurrent to the examination situation and not as prospective or retrospective. Added to this, the items failed to discriminate between verbalizations related to the activity in progress or their outcomes. From the data, it can be deduced that the items evaluate teachers' verbalizations whose content provides information regarding the distinguishable achievement only by their valence (Pekrun, 2018).

In contrast, a large number of items were designed with the intention of arriving at a brief instrument with the best items that facilitate and make their application more practical. After identifying the items with the best factor loadings, two scales of eight items each with very good internal consistency were retained.

Teachers' achievement verbalizations can affect control and value appraisals and, through these, activate achievement emotions and academic performance (Goetz et al., 2018). This research provided favorable evidence on the relationship of the scores of each scale with theoretically related variables (Apto et al., 2017; Pekrun et al., 2023; Putwain et al., 2017). Low correlations with achievement emotions were obtained, although the association seems to be greater between verbalizations and emotions of equal valence. There were also positive and negative relationships between the positive and negative verbalizations, respectively, with the academic average. The magnitudes of these relationships, although low, seem consistent with the idea that control-value appraisals would be mediating this association (Pekrun, 2018, 2021).

Limitations and further studies

Although the instrument developed presents some good initial psychometric properties, the results should be taken with caution and further studies should be carried out. First, the items express verbalizations made during oral exams and fail to capture the verbalizations that are anticipated several days before, for example, in a class situation or that are even carried out a few days later. This distinction, not achieved with our development of items, could have differential effects not only on achievement emotions, but also on, for example, behavioral avoidance in oral exams (Furlan & Sánchez-Rosas, 2018). Verbalizations before or after an exam would affect the postponement of the evaluation instances, while those that are said during the exam would have a greater impact on inhibition during the oral exam.

Conversely, the clear gender bias in the sample has not made it possible to verify whether the measurements remain invariant based on sex. This would be important to analyze since, if achievement emotions related to negative outcomes seem to be more frequent in women (Reilly & Sánchez-Rosas, 2019; Sánchez-Rosas, 2015), verbalizations, considering their environmental antecedents, could also vary depending on this categorical variable.

Finally, although the constructed items clearly refer to verbal expressions on achievement, it would be convenient to consider an analysis of convergence or divergence through correlation with instruments that assess feedback or fear appeals (Putwain et al., 2017, 2023), instructional teaching quality in class (Becker et al., 2014; Lazarides & Buchholz, 2019; Narciss et al., 2022; Sánchez-Rosas et al., 2016), teacher support (Apto et al., 2017; Lei et al., 2018) or the inclusion of non-verbal behaviors (Derakhshan et al., 2023; Guo et al., 2022; Juma et al., 2022; Puertas-Molero et al., 2022).

Practical implications

Through the TAVQ, the measurement of teachers' verbalizations during oral exams makes it possible to broaden and enrich the functional analysis of emotional dysregulation problems and avoidance behaviors in evaluative contexts (Furlan & Sánchez-Rosas, 2018). The messages related to achievement during an oral exam offer feedback on the control and value appraisals that precede the emotional responses of the people evaluated and that later activate their coping behaviors (Pekrun, 2018, 2021). The information processing related to performance during the exam is part of executive control tasks when goal-directed behaviors are implemented (Zeidner & Matthews, 2005). In this way, the information provided by teachers will be processed by each student according to their beliefs and appraisals and will lead to behaviors that tend to regulate their emotional state, using the set of strategies that can be accessed (Rojas-Torres et al., 2022). For this reason, it is valuable to have an evaluation tool that allows one to reflect on the nature and effects of verbal messages, noting their relevance in students' performances in oral exams.

Conclusions

An instrument is provided with two scales that evaluate positive and negative teachers' achievement verbalizations in oral exams with evidence of validity and reliability. In addition to the optimal reliability values, the relationship of the scores of each scale with the achievement emotions and academic performance is demonstrated. In short, we count with a useful instrument for the assessment of verbal expressions, phrases and comments about students' achievement that a teacher emits in the presence of one or more students during an oral exam.

References

Apto, M. S., Pesqueira, N. G., Vilca, B., & Sánchez-Rosas, J. (2017). Efecto y contribución explicativa del apoyo e inhibición, interacción-ilustración, vergüenza, disfrute y amenazas a la evitación de la búsqueda de ayuda académica. *Anuario de Investigaciones de la Facultad de Psicología, 3*(1), 247-263. https://revistas.unc.edu.ar/index.php/aifp/article/view/18670

Awad-Igbaria, Y., Maaravi-Hesseg, R., Admon, R., & Karni, A. (2022). Only tomorrow: Delayed effects of teachers attitude on motor skill learning. *Learning and Instruction, 82*. https://doi.org/10.1016/j.learninstruc.2022.101681

Becker, E. S., Goetz, T., Morger, V., & Ranellucci, J. (2014). The importance of teachers' emotions and instructional behavior for their students' emotions – An experience sampling analysis. *Teaching and Teacher Education: An International Journal of Research and Studies, 43*(1), 15-26. https://doi.org/10.1016/j.tate.2014.05.002

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

Browne, M. W., & Cudeck, R. (1993). Alternative ways of assessing model fit. In K. A. Bollen & J. S. Long (Eds.), *Testing structural equation models* (pp. 136-162). Sage.

Burić, I. (2015). The role of social factors in shaping students' test emotions: A mediation analysis of cognitive appraisals. *Social Psychology of Education, 18*(4), 785-809. https://doi.org/10.1007/s11218-015-9307-9

Derakhshan, A., Zhang , L. J., & Zhaleh, K. (2023). The effects of instructor clarity and non-verbal immediacy on Chinese and Iranian EFL students' affective learning: The mediating role of instructor understanding. *Studies in Second Language Learning and Teaching, 13*(1), 71-100. https://doi.org/10.14746/ssllt.31733

Dewaele, J.-M., Magdalena, A. F., & Saito, K. (2019). The effect of perception of teacher characteristics on Spanish EFL learners' anxiety and enjoyment. *The Modern Language Journal, 103*(2), 412-427. https://doi.org/10.1111/modl.12555

Dewaele, J.-M., Witney, J., Saito, K., & Dewaele, L. (2018). Foreign language enjoyment and anxiety: The effect of teacher and learner variables. *Language Teaching Research, 22*(6), 676-697. https://doi.org/10.1177/1362168817692161

DiStefano, C., Liu, J., Jiang, N., & Shi, D. (2018). Examination of the weighted root mean square residual: Evidence for trustworthiness? *Structural Equation Modeling: A Multidisciplinary Journal, 25*(3), 453-466. https://doi.org/10.1080/10705511.2017.1390394

Dominguez-Lara, S. (2018). Propuesta de puntos de corte para cargas factoriales: Una perspectiva de fiabilidad de constructo. *Enfermería Clínica, 28*(6), 401-402. https://doi.org/10.1016/j.enfcli.2018.06.002

Finney, S. J., & DiStefano, C. (2006). Non-normal and categorical data in structural equation modeling. In G. R. Hancock & R. O. Mueller (Eds.), *Structural Equation Modeling: A Second Course* (pp. 269-314). Information Age.

Frenzel, A. C., Daniels, L., & Burić, I. (2021). Teacher emotions in the classroom and their implications for students. *Educational Psychologist, 56*(4), 250-264. https://doi.org/10.1080/00461520.2021.1985501

Furlan, L., & Sánchez-Rosas, J. (2018). Evidencias de validez y confiabilidad de una Escala de Evitación Conductual en Exámenes Orales en estudiantes universitarios. *Ansiedad y Estrés, 24*(2-3), 90-98. https://doi.org/10.1016/j.anyes.2018.05.001

Gardner, D. E., & Giordano, A. N. (2023). The challenges and value of undergraduate oral exams in the physical chemistry classroom: A useful tool in the assessment toolbox. *Journal of Chemical Education, 100*(5), 1705-1709. https://doi.org/10.1021/acs.jchemed.3c00011

Goetz, T., Lipnevich, A. A., Krannich, M., & Gogol, K. (2018). Performance feedback and emotions. In A. A. Lipnevich & J. K. Smith (Eds.), *The Cambridge Handbook of Instructional Feedback* (pp. 554-574). Cambridge University Press. https://doi.org/10.1017/9781316832134.027

Guo, H., Gao, W., & Shen, Y. (2022). L2 Enjoyment of English as a foreign language students: Does teacher verbal and non-verbal immediacy matter? *Frontiers in Psychology, 13*. https://doi.org/10.3389/fpsyg.2022.897698

Hazen, H. (2020). Use of oral examinations to assess student learning in the social sciences. *Journal of Geography in Higher Education, 44*(4), 592-607. https://doi.org/10.1080/03098265.2020.1773418

Hunsley, J., & Marsh, E. J. (2008). Developing criteria for evidence-based assessment: An introduction to assessments that work. In J. Hunsley & E. J. Marsh (Eds.), *A guide to assessments that work* (pp. 3-14). Oxford University Press.

Juma, O., Husiyin, M., Akhat, A., & Habibulla, I. (2022). Students' classroom silence and hopelessness: The impact of teachers' immediacy on mainstream education. *Frontiers in Psychology, 12*. https://doi.org/10.3389/fpsyg.2021.819821

Lazarides, R., & Buchholz, J. (2019). Student-perceived teaching quality: How is it related to different achievement emotions in mathematics classrooms? *Learning and Instruction, 61*, 45-59. https://doi.org/10.1016/j.learninstruc.2019.01.001

Lei, H., Cui, Y., & Chiu, M. M. (2018). The relationship between teacher support and students' academic emotions: A meta-analysis. *Frontiers in Psychology,* 8. https://doi.org/10.3389/fpsyg.2017.02288

McDonald, R. P., & Ho, M.-H. R. (2002). Principles and practice in reporting structural equation analyses. *Psychological Methods, 7*(1), 64-82. https://doi.org/10.1037/1082-989X.7.1.64

Muthén, L. K., & Muthén, B. O. (2012). Mplus User's Guide. Muthén & Muthén Press.

Narciss, S., Prescher, C., Khalifah, L., & Körndle, H. (2022). Providing external feedback and prompting the generation of internal feedback fosters achievement, strategies and motivation in concept learning. *Learning and Instruction, 82*. https://doi.org/10.1016/j.learninstruc.2022.101658

Pekrun, R. (2018). Control-value theory: A social-cognitive approach to achievement emotions. In G. A. D. Liem & D. M. McInerney (Eds.), *Big theories revisited 2: A volume of research on sociocultural influences on motivation and learning* (pp. 162-190). Information Age Publishing.

Pekrun, R. (2021). Self-appraisals and emotions: A generalized control-value approach. In T. Dicke, F. Guay, H. W. Marsh, R. G. Craven, & D. M. McInerney (Eds.), *Self - A multidisciplinary concept* (pp. 1-30). Information Age.

Pekrun, R., Marsh, H. W., Elliot, A. J., Stockinger, K., Perry, R. P., Vogl, E., Goetz, T., van Tilburg, W. A. P., Lüdtke, O., & Vispoel, W. P. (2023). A three-dimensional taxonomy of achievement emotions. *Journal of Personality and Social Psychology, 124*(1), 145-178. https://doi.org/10.1037/pspp0000448

Pérez, E., Medrano, L., & Sánchez-Rosas, J. (2013). El Path Analysis: Conceptos básicos y ejemplos de aplicación. *Revista de la Asociación Argentina de Ciencias del Comportamiento, 5*(1), 52-66. https://revistas.unc.edu.ar/index.php/racc

Ponterotto, J., & Charter, R. A. (2009). Statistical extensions of Ponterotto and Ruckdeschel's (2007) Reliability Matrix for estimating the adequacy of internal consistency coefficients. *Perceptual and Motor Skills, 108*(3), 878-886. https://doi.org/10.2466/PMS.108.3.878-886

Puertas-Molero, P., Zurita-Ortega, F., González-Valero, G., & Ortega-Martín, J. L. (2022). Design and validation of the Non-Verbal Immediacy Scale (NVIS) for the evaluation of non-verbal language in university professors. *International Journal of Environmental Research and Public Health, 19*(3), 1159. http://dx.doi.org/10.3390/ijerph19031159

Putwain, D. W., Nakhla, G., Liversidge, A., Nicholson, L. J., Porter, B., & Reece, M. (2017). Teachers use of fear appeals prior to a high-stakes examination: Is frequency linked to perceived student engagement and how do students respond? *Teaching and Teacher Education, 61*, 73-83. https://doi.org/10.1016/j.tate.2016.10.003

Putwain, D. W., Nicholson, L. J., & Kutuk, G. (2023). Warning students of the consequences of examination failure: An effective strategy for promoting student engagement? *Journal of Educational Psychology, 115*(1), 36-54. https://doi.org/10.1037/edu0000741

Putwain, D. W., Pekrun, R., Rainbird, E., & Roberts, C. (2022). Cognitive-behavioural intervention for test anxiety: Could teachers deliver the STEPS Program and what training would they require? In L. R. V. Gonzaga, L. L. Dellazzana-Zanon, & A. M. Becker da Silva (Eds.), *Handbook of Stress and Academic Anxiety* (pp. 381-399). https://doi.org/10.1007/978-3-031-12737-3_25

Raccanello, D., Hall, R., & Burro, R. (2018). Salience of primary and secondary school students' achievement emotions and perceived antecedents: Interviews on literacy and mathematics domains. *Learning and Individual Differences, 65*, 65-79. https://doi.org/10.1016/j.lindif.2018.05.015

Reilly, P., & Sánchez-Rosas, J. (2019). The achievement emotions of English language learners in Mexico. *Electronic Journal of Foreign Language Teaching, 16*(1), 34-48. http://e-flt.nus.edu.sg/v16n12019

Reilly, P., & Sánchez-Rosas, J. (2021). Achievement emotions and gender differences associated with second language testing. *International Journal of Instruction, 14*(4), 825-840. https://doi.org/10.29333/iji.2021.14447a

Rodriguez, A., Reise, S. P., & Haviland, M. G. (2016). Evaluating bifactor models: Calculating and interpreting statistical indices. *Psychological Methods, 21*(2), 137-150. https://doi.org/10.1037/met0000045

Rojas-Torres, L., Furlan, L. A., Sánchez-Rosas, J., & Rojas-Rojas, G. (2022). Construcción y validación de la Escala de Afrontamiento de la Ansiedad durante los Exámenes. *Ansiedad y Estrés, 28*(2), 81-90. https://doi.org/10.5093/anyes2022a9

Sánchez-Rosas, J. (2015). The Achievement Emotions Questionnaire-Argentine (AEQ-AR): Internal and external validity, reliability, gender differences and norm-referenced interpretation of test scores. *Evaluar, 15*(1), 41-74. https://doi.org/10.35670/1667-4545.v15.n1.14908

Sánchez-Rosas, J. (2016). Avances preliminares en la identificación de causas docentes de la ansiedad, vergüenza y disfrute en exámenes orales [Comunicaciones libres]. *Revista Argentina de Ciencias del Comportamiento, Suplemento I (Agosto)*, 108-109. https://revistas.unc.edu.ar/index.php/racc

Sánchez-Rosas, J., Correa, P., & Díaz, I. (2019). Revisión de las intervenciones que mejoran la utilidad percibida del aprendizaje de los estudiantes. *Revista Digital de Investigación en Docencia Universitaria, 13*(2), 41-52. https://revistas.upc.edu.pe/index.php/docencia

Sánchez-Rosas, J., & Esquivel, S. (2016). Instructional teaching quality, task value, self-efficacy, and boredom: A model of attention in class. *Revista de Psicología, 25*(2), 1-20. https://doi.org/10.5354/0719-0581.2016.44966

Sánchez-Rosas, J., & Furlan, L. A. (2017). Achievement emotions and achievement goals in support of the convergent, divergent and criterion validity of the Spanish-Cognitive Test Anxiety Scale. *International Journal of Educational Psychology, 6*(1), 67-92. https://doi.org/10.17583/ijep.2017.2268

Sánchez-Rosas, J., Takaya, P. B., & Molinari, A. V. (2016). The role of teacher behavior, motivation and emotion in predicting academic social participation in class. *Pensando Psicología, 12*(19), 39-53. https://doi.org/10.16925/pe.v12i19.1327

Theobold, A. S. (2021). Oral exams: A more meaningful assessment of students' understanding. *Journal of Statistics and Data Science Education, 29*(2), 156-159. https://doi.org/10.1080/26939169.2021.1914527

Ventura-León, J., Caycho-Rodríguez, T., Sánchez-Villena, A. R., Peña-Calero, B. N., & Sánchez-Rosas, J. (2022). Academic inspiration: Development and validation of an instrument in higher education. *Electronic Journal of Research in Education Psychology, 20*(58), 635-660. https://doi.org/10.25115/ejrep.v20i58.5599

Westphal, A., Kretschmann, J., Gronostaj, A., & Vock, M. (2018). More enjoyment, less anxiety and boredom: How achievement emotions relate to academic self-concept and teachers' diagnostic skills. *Learning and Individual Differences, 62*, 108-117. https://doi.org/10.1016/j.lindif.2018.01.016

Yang, Y., Gao, Z., & Han, Y. (2021). Exploring Chinese EFL learners' achievement emotions and their antecedents in an online English learning environment. *Frontiers in Psychology, 12*. https://doi.org/10.3389/fpsyg.2021.722622

Zeidner, M., & Matthews, G. (2005). Evaluation Anxiety: Current Theory and Research. In A. J. Elliot & C. S. Dweck (Eds.), *Handbook of competence and motivation* (pp. 141-163). Guilford Publications.