Psicothema

Psicothema 2018, Vol. 30, No. 4, 442-448
doi: 10.7334/psicothema2018.113

ISSN 0214 - 9915 CODEN PSOTEG
Copyright © 2018 Psicothema
www.psicothema.com

Seven methods to determine the dimensionality of tests: application to the General Self-Efficacy Scale in twenty-six countries

Greibin Villegas Barahona[1,2], Nerea González García[1], Ana Belén Sánchez-García[1], Mercedes Sánchez Barba[1], and María Purificación Galindo-Villardón[1]

[1] Universidad de Salamanca and [2] Universidad Estatal a Distancia Costa Rica

Abstract

Background: One of the most important concepts within Cognitive Social Theory as framed by Bandura is the perceived self-efficacy; this concept became widespread in 1981 when Mathias Jerusalem and Ralf Schwarzer, using 10 items, established a one-dimensional and universal construct of this scale. The main purpose of this study is to show that the General Self-Efficacy Scale (GSE) is not a one-dimensional and universal construct, as is currently assumed. **Method:** The data from 19,719 people from 26 countries were analyzed. In order to identify and understand invariance we applied seven multivariate statistical techniques. **Results:** The findings suggest the existence of a multidimensional structure and differential item functioning by country. Insofar as there is differential item functioning by country and it is not possible to universalize it, and there are several items on the scale that statistically constitute additional factors. The results confirm that the self-efficacy construct is neither universal nor unidimensional. **Conclusions:** A psychometric instrument must be valued and used with great care; the one in question is being used in a generalized way.

Keywords: self-efficacy, Item Response Theory, dimensionality, cross-cultural comparisons, invariance.

Resumen

Siete métodos para evaluar la dimensionalidad de los test: aplicación a la General Self-Efficacy Scale en veintiséis países. **Antecedentes:** uno de los conceptos más importantes en la Teoría Social Cognitiva desarrollada por Bandura es la auto-eficacia percibida. Este concepto ha sido generalizado en 1981 por Mathias Jerusalem and Ralf Schwarzer con una escala de 10 ítems, quienes establecieron que esta escala es un constructo unidimensional y universal. El objetivo principal de este trabajo es demostrar que la Escala General de Autoeficacia (GSE) no es un constructo unidimensional ni universal, como actualmente se asume. **Método:** los datos analizados corresponden a 19.719 personas de 26 países. Con el fin de identificar y entender la invariancia hemos utilizado siete técnicas estadísticas multivariantes. **Resultados:** los hallazgos sugieren la existencia de una estructura multidimensional y un funcionamiento diferencial por país. En la medida que haya funcionamiento diferencial por país, no es posible universalizar el constructo. También existen varios ítems de la escala que constituyen factores adicionales. El resultado confirma que el constructo auto-eficacia no es universal ni unidimensional. **Conclusiones:** un instrumento psicométrico debe ser evaluado y usado con extremo cuidado, la escala GSE analizada está siendo utilizada de manera general.

Palabras clave: autoeficacia, Teoría Respuesta al Item, dimensionalidad, comparaciones culturales, invariancia.

The Multifactorial Nature of the Self-efficacy Construct

One of the most important concepts within the Cognitive Social Theory framed by Bandura is perceived self-efficacy. Bandura, (1977, 1978, 1982, 1986) establishes a causal relationship between each person's perceived self-efficacy and the effort he or she expends to face the challenges and goals regulating their different processes (e.g, cognitive, affective, motivational and action), concluding that high levels of perceived self-efficacy are projected through the existence of superior performance achievements, as well as reduction of stress and depression (Bandura, 1999).

Starting with Bandura's theory (2012), the concept of perceived self-efficacy of a multifactorial nature is outlined relative to a single domain or specific task, which must be evaluated in terms of capacity by reference to that domain. For this reason, the items should be formulated in those terms. Carrasco & Del Barrio (2002) confirmed in their research and insist that the construct of self-efficacy is linked to exact and specific areas, in the face of the consideration of a single global trait.

Bandura (2012) points out the misuse that has been made of the theoretical conceptualization and the way self-efficacy is measured and states that the domains of complex activity need to be evaluated in a multi-dimensional manner, particularly the different types of self-efficacy which operate together. Bandura points out that it is a mistake to characterize self-efficacy in a narrow field, using self-efficacy measures of a generalized nature and without contextualized content.

However, other authors such as Scholz et al. (2002) and Schwarzer & Jerusalem (1995) have projected the concept of self-

Received: April 24, 2018 • Accepted: August 27, 2018
Corresponding author: Greibin Villegas Barahona
Universidad de Salamanca
Universidad Estatal a Distancia Costa Rica
37004 Salamanca (Spain)
e-mail: gvillegas@uned.ac.cr

efficacy with a general orientation. This is possibly applicable to different beliefs of the individual regarding their expectations of confidence to face diverse situations that could potentially generate stress. Accordingly, this orientation is reflected in the General Self-Efficacy Scale (GSE).

In the last century, Thurstone (1931) properly develops the technique of Factorial Analysis in designing of psychometric instruments; whose origin can be traced back to Spearman's work (1904). The simple structural principle states that any psychometric instrument must be explained by one or only some of the latent factors. In order to achieve a lower number of constructs, a large part of the variability explained is sacrificed, it is important not to ignore this unexplained information; since it could be the cause behind why the results can be invalidated due to a differential operation.

Most psychometric instruments have been developed with the basic foundations of classical test theory (CTT) proposed by Spearman at the beginning of the 20th century (1904). However, the CTT approach presents limitations mainly due to the lack of invariance of the measurements with respect to the instrument used (Muñiz, 2010) and regarding the sample used in the research (interculturality), which limits the generalization to different populations (Brown, Harris, O'Quin, & Lane, 2015). Another limitation is the reliability of an estimation of the scores in the classical approach. It is assumed that the instruments measure all the people evaluated with certain reliability, but this is not the case. It is clear that differences between populations (e.g. Orient-Occident, Europe-Latin America and Asia-Africa) affect the reliability of an estimation of the scores, possibly because the population's cultural differences are not included in the low variance percentages by means of a single one-dimensional factor.

For example, the German version of the self-efficacy construct has been universalized as a one-dimensional latent construct (Scholz et al., 2002). However, when the variance explained from the one-dimensional model is close to 43%, it is probable that other factors are hidden in the remaining 57% of the unexplained variance. This unexplained variance can contain relevant specific information not captured in the shared latent structure; which does not visualize the differential item functioning (DIF) between countries, causing a lack of invariance. For this reason, it is appropriate to complement it with other statistical techniques such as the Multigroup Confirmatory Factor Analysis to evaluate the Cross-Cultural aspect, in order to identify and understand the invariance.

Taking into account the theoretical framework described above, the main objective is to examine the dimensionality of the German version of the GSE construct. The second objective is to analyze the differential item functioning between cultures to assess the universality of this construct. To achieve these objectives, a multivariate statistical analysis is performed with the following techniques: Factor Analysis (FA), Principal Components Analysis (PCA), Sparse Principal Components Analysis (Sparse PCA), Dual Statis, Item Response Theory (IRT), Differential Item Functioning (DIF) and Multi-group Confirmatory Factor Analysis to evaluate the cross-cultural aspect (MGCFA). The first three techniques allow assessing the existence of common factors; the fourth builds a consensus matrix from the matrices of variance and covariance of the 26 countries representing the similarities and differences among in them.

Methods

Participants

The database used is composed of the accumulation of different studies carried out on students, adults, police officer candidates, immigrants, air force and armed forces soldiers, parents, educators, teachers, and nurses, among others. The characteristics of these populations are found in Scholz et al. (2002), who account for the heterogeneity of the sample in activity and age. The database available at: http://userpage.fu-berlin.de/~health/selfscal.htm. It includes N=19,896 records from 26 countries (Table 1). In the case of Switzerland, there are 177 individuals missing all responses, therefore N=19,719. In the database, there are 165 cases with missing values that received an imputation. At the moment Swedish, Bulgarian, Armenian, Urdu (Pakistani), Slovenian, Serbian, and Brazilian participants have been added to this investigation, but in the statistical analysis of this article, they will not be taken into account, since said data were not yet available.

Instrument

The GSE scale of Matthias Jerusalem and Ralf Schwarzer (1995) consists of 10 items, evaluated with a Likert scale of four points, according to the following categorization: 1 "Not at all true", 2 "Hardly true", 3 "Moderately true" y 4 "Exactly true". This instrument was translated into 28 languages (Schwarzer & Jerusalem, 1995; Scholz et al., 2002) by bilingual native speakers; subsequently, the "group consensus model" with several bilingual translators was applied. The procedure included back translations and group discussions, "Since the goal was to achieve cultural-sensitive adaptations of the construct, rather than mere literal translations, the translation sought a thorough understanding of the general self-efficacy construct" (Scholz et al., 2002). One of the aspects favoring this scale is that its items have been designed in a positive direction, which reduces the probability of response bias according to Suárez-Alvarez et al. (2018).

Given that this scale is used for the calculation of the latent variable and it is answered in a self-administered manner, in principle this fact could affect the validity of the findings discovered; a situation faced by Scholz et al. (2002). For the purpose of obtaining evidence on the validity of the internal structure of the instruments, we suggest reviewing Sireci and Padilla (2014) and The Standards for Educational and Psychological Testing Developed jointly by the American Educational Research Association (AERA), American Psychological Association (APA) and National Council on Measurement in Education (NCME).

Procedure

An analysis was carried out with the same database used by its authors, all the details about the data collection, ethics issues followed in the search of the data can also be seen in Scholz et al. (2002), regarding the way they created the database by accumulation of data on a country-by-country basis.

On the other hand, as indicated by Lorenzo-Seva & Ferrando (2011) "Polychoric correlation is advised when the univariate distributions of ordinal items are asymmetric or with an excess of kurtosis. If both indices are lower than one in absolute value, then Pearson correlation is advised" (p. 12). In this case, the 91.0% of the asymmetric and kurtosis indicators are lower than one in absolute value. We use Maximum likelihood estimation (ML) in FA and PCA. In this analysis we use the DIF detection method based on the Item Response Theory (IRT) under the Graded Response Model (GRM) of Samejima (1969). To this end we used the MULTILOG program and parameter estimation with maximum marginal likelihood (version 7.03; Thissen, 2003).

Data analysis

In order to demonstrate the importance of unexplained variability and the lack of invariance, the GSE designed in 1981 by Mathias Jerusalem and Ralf Schwarzer with 10 items and adapted to 28 languages is taken as an example. To verify this, we performed a secondary analysis of data (n= 19,719) using seven different multivariate statistical techniques: FA, PCA, Sparse PCA, Dual Statis, IRT, FD, and MGCFA.

FA (Spearman, 1904) is one of the most commonly used multivariate statistical techniques in order to study the latent factors underlying the relationships between variables. PCA with FA are proposed as dimension reduction techniques, to be determined with the same criteria used by the authors (Scholz et al., 2002), when self-efficacy corresponds to a single factor in all countries. The Sparse PCA looks for a useful, simple and interpretable factorial solution. This is the reason a methodological line associated with PCA has been developed from the rotation methods (Jolliffe, 2002), which evolves to techniques of variable selection such as Sparse Principal Components Analysis (Ning-min & Jing, 2015). Among the existing algorithms, the formulation proposed by Zou, Hastie, and Tibshirani is used (2006) available in free software R (Version 1.1).

Dual Statis is an extension of the principal components analysis proposed to analyze more than one data matrix simultaneously (Abdi, Williams, Valentin, & Bennani, 2012). With this technique, it is possible to compare the matrices of variance and covariance of each country with the consensus matrix obtained for the entire database with the weighting of all of them and thus determine the existence of similarities and differences in the GSE scale in the different countries.

IRT analyzes the behavior of the GSE scale at the level of each item and at the combined level of the whole scale (Hambleton, 1996). While the DIF detects if an item works similarly or differently for one country than for another and if the instrument as a whole is the one that shows different statistical properties in each group (Ackerman, 1992; Goldstein, 1980; Gómez-Benito, Balluerka, González, Widaman, & Padilla, 2017; Lord, 1980; Samejima, 1969; Raju, 1999). The criterion of Non-compensatory Differential Item Functioning (NCDIF) was used, which is an item level statistic that reflects the differences in the scores in pairs of countries. There are critical values according to the number of categorical answers for the significance of the NCDIF for an item; in this particular case, the critical value is 0.054 for a $\alpha=0.01$ (Rajú, 1999).

The MGCFA technique determines if the items of a measurement instrument have similar patterns of response in all the countries observed. If statistical equality is found in the results then it is possible to compare the scores of measurement instruments and it can be said that the samples come from the same population (Brown et al., 2015), confirming the universality of the instrument. With this technique, we evaluated the model considering the Mean and the Covariance Structures (MACS), while some others only compare the covariance structures (COVS).

To evaluate the fit of the model, different recommended indicators from the literature were considered. (Byrne, 2001; Cheung & Rensvold, 2002; Hu & Bentler, 1999). In order to be an acceptable model, these indices should be$\chi^2/gl<3.8$, *RMSEA* (Root Mean Square Error of Approximation) < 0.08, *SRMR* (Standarized Root Mean Square) *≈0.06* y *CFI* (Comparative Fit Index) *>0.9*. In the same way, in order for the model to be good, these values must be $\chi^2/gl<3.0$, *RMSEA* < 0.05, *SRMR <0.06* y *CFI >0.95*.

Results

In order to describe the possible unidimensionality of each country, a factorial analysis was applied to determine the number of components to be retained and the explained variance when considering the trait as one-dimensional. First, the Bartlett sphericity test (all significant) and the Kaiser-Meyer-Olkin Measure of sampling adequacy (KMO) were performed for the entire sample and for each country, the results show optimal values for assuming that it is possible find a latent structure in the data matrix (Table 1 and Table 4).

Table 1 shows that of the 26 countries analyzed, 13 of them managed to obtain a single principal component with an average explained variance of 45.0%. This first result shows that half of the countries have more than one main component. When extracting the components in the whole sample (n = 19,719) a single factor is obtained with 43.7% leaving 56.3% without explanation (Table 1).

When comparing the results of the FA, the PCA and the SPCA (Table 2), it is clearly shown that item 1 and item 6: "I can solve most problems if I invest the necessary effort" form a second factor and that item 2 is a single factor by itself. When stablishing a one-dimensional construct, information on those items is being lost, since the first latent dimension does not collect the respective information from them, as it is found in later dimensions. It is known that latent constructs should be constructed with at least two items; however, in this case, the idea is to verify whether item 2 has special features that make it a component by itself.

In the case of three component extraction, in the first component it is observed that the items are heavily loaded: " I am confident that I could deal efficiently with unexpected events"; 5: "Thanks to my resourcefulness, I know how to handle unforeseen situations"; 7: "I can remain calm when facing difficulties because I can rely on my coping abilities"; 8: "When I am confronted with a problem, I can usually find several solutions"; 9: "If I am in trouble, I can usually think of something to do"; and those of less loading in this component are items 3 and 10: "No matter what comes my way, I´m usually able to handle it".

Where it can be seen that items 1, 2 and 3 are those that present different characteristics. A special analysis requires the result of Sparse PCA, a technique that assigns very small factor loads, close or equal to zero, in order to truly extract the item that corresponds to each component. For this reason, in the case of item 2 in the three-component modality, Component 2 is fully loaded, leaving the factorial loads to the other components equal to zero. This confirms that the low loads of items 1, 2 and 3 in Component 1 are due to item 1 and item 2 loading strongly in a second and third component. SPCA improves the interpretation of results, to the

Table 1
Sample and variance explained with PCA by country, 2017

N.	Countries	n	% of n	PCA[a]	Variance explained by PCA			% Variance explained
					PC 1	PC 2	PC 3	
1	Indonesia	536	2.7%	1	36.0%			36.0%
2	Germany	7100	36.0%	1	37.4%			37.4%
3	Costa Rica	943	4.8%	1	39.1%			39.1%
4	Iran	802	4.1%	1	41.5%			41.5%
5	Russia	495	2.5%	1	42.8%			42.8%
6	Poland	690	3.5%	1	43.3%			43.3%
7	USA	1594	8.1%	1	46.5%			46.5%
8	Korea	147	0.7%	1	47.7%			47.7%
9	Denmark	153	0.8%	1	47.8%			47.8%
10	Canada	367	1.9%	1	48.4%			48.4%
11	Hungary	158	0.8%	1	48.9%			48.9%
12	Great Britain	447	2.3%	1	49.5%			49.5%
13	Japan	430	2.2%	1	56.1%			56.1%
14	Portugal	544	2.8%	2	24.1%	19.3%		43.4%
15	Peru	994	5.0%	2	28.5%	17.8%		46.3%
16	Switzerland	599	3.0%	2	25.7%	21.1%		46.8%
17	Italy	144	0.7%	2	25.0%	22.0%		47.0%
18	Syria	264	1.3%	2	25.0%	23.2%		48.2%
19	Netherlands	911	4.6%	2	33.0%	20.4%		53.3%
20	Spain	399	2.0%	2	36.4%	17.5%		53.9%
21	Hong Kong	1067	5.4%	2	34.7%	19.2%		54.0%
22	Finland	159	0.8%	2	28.5%	26.5%		55.0%
23	Belgium	175	0.9%	2	39.9%	16.6%		56.5%
24	India	398	2.0%	3	21.4%	16.5%	14.4%	52.3%
25	Greece	100	0.5%	3	21.3%	20.4%	17.9%	59.6%
26	France	103	0.5%	3	26.3%	20.5%	17.6%	64.4%
	General	19719	100.0%	1	43.7%			43.7%

[a] Number of PCA Principal Component Analyses Extracted

Table 2
Factorial loading with different methods, 2017

Item	FA 1/			PCA 1/			SPARSE PCA2/			SPARSE PCA3/		
	1	2	3	1	2	3	1	2	3	1	2	3
Self1	0.18	0.30	0.45	0.12	0.88	0.27	0.00	-0.88	0.00	0.00	-0.90	0.00
Self2	0.19	0.56	0.14	0.18	0.13	0.89	0.00	0.00	0.99	0.00	-0.11	0.92
Self3	0.41	0.35	0.22	0.56	0.06	0.37	-0.32	0.00	0.00	-0.28	0.00	0.17
Self4	0.64	0.22	0.25	0.76	0.12	0.09	-0.43	0.03	0.00	-0.43	0.09	0.00
Self5	0.63	0.26	0.26	0.74	0.15	0.15	-0.42	0.00	0.00	-0.42	0.06	0.00
Self6	0.35	0.12	0.57	0.51	0.61	-0.10	-0.16	-0.47	-0.05	-0.26	-0.40	-0.28
Self7	0.50	0.28	0.33	0.70	0.19	0.12	-0.40	0.00	0.00	-0.39	0.00	0.00
Self8	0.38	0.44	0.28	0.54	0.17	0.42	-0.32	0.00	0.10	-0.29	0.00	0.21
Self9	0.44	0.31	0.36	0.62	0.30	0.16	-0.37	0.00	0.00	-0.37	0.00	0.00
Self10	0.43	0.37	0.31	0.59	0.23	0.28	-0.36	0.00	0.00	-0.35	0.00	0.00
%Ev	19.6	11.4	11.4	32.6	14.0	13.4	37.7	8.6	8.3	38.2	8.6	8.6
% Cum		42.38			59.99			54.60			55.40	

1/ Varimax rotation
2/ spca(x=cgse, K=3, para = c(8,3,3, type = "Gram", sparse = "vamum", trace = TRUE)
3/ spca(x=cgse, K=3, para = c(8,5,5, type = "Gram", sparse = "vamum", trace = TRUE)
%EV Explained variance
%Cum Cumulative variance

detriment of the percentage of variability explained (Zou et al., 2006). This allows clarifying the location of the items that belong to a certain factor.

Regarding the Statis Dual analysis, the first characteristic vector of each of the matrices is used and a single vector is formed that is represented by the first axis of the matrix commitment. Table 3 shows how there are differences in the matrices of variance and covariance of the 26 countries. On the other hand, in countries such as Japan and India, as they move away from the abscissa axis, the similarity to that commitment decreases, which consists of a weighted matrix.

In addition, the value of the Hilbert-Schmidt standard (NS Norms[2]) confirms the information established above, since, the higher the standard, that country contributes in a better way to the construction of the commitment. On the other hand, the differences found in the cosines to the square of the angles, explain the differences found between countries; since, at a lower angle with the axis of the abscissa, the variance and covariance matrix of a particular country tends to be similar to the commitment matrix.

If the GSE scale offered the same results among the countries, all the vectors that represent the variance and covariance matrices should have an angle close to zero with the tendency to parallel the abscissa axis.

Table 3
Weights, NS Norms^2 and Cos^2 between countries and compromise, 2017

Countries	Weights[1/]	(NS Norms^2)[2/]	(Cos^2)[3/]
Japan	* 0.24	* 0.34	* 0.95
Great Britain	* 0.22	* 0.28	* 0.92
Canada	* 0.22	* 0.27	* 0.91
USA	* 0.21	* 0.25	* 0.91
Hungary	* 0.22	* 0.27	* 0.89
Poland	* 0.20	0.23	* 0.89
Finland	* 0.20	0.24	* 0.88
Denmark	* 0.21	* 0.26	* 0.88
Netherlands	0.20	0.23	* 0.88
Korea	* 0.21	* 0.26	* 0.88
Hong Kong	0.20	0.23	* 0.88
Belgium	0.20	* 0.24	* 0.87
Spain	0.20	0.23	* 0.87
Iran	0.20	0.21	0.86
France	0.19	0.22	0.84
Russia	0.20	0.22	0.84
Costa Rica	* 0.19	* 0.20	* 0.82
Germany	* 0.19	* 0.18	* 0.82
Switzerland	* 0.18	* 0.18	* 0.81
Italy	* 0.18	* 0.18	* 0.81
Indonesia	* 0.18	* 0.18	* 0.80
Peru	* 0.18	* 0.18	* 0.80
Greece	* 0.17	* 0.18	* 0.79
Syria	* 0.17	* 0.18	* 0.76
Portugal	* 0.18	* 0.16	* 0.75
India	* 0.16	* 0.15	* 0.70
M	0.20	0.22	0.85
SD	0.02	0.04	0.06
Min	0.16	0.15	0.70
Max	0.24	0.34	0.95
95% CI LL	0.19	0.205	0.82
UL	0.20	0.238	0.88
Kolmogorov Test	p =.200	p =.200	p =.200

1/ "Weights" indicates the weight acquired by each matrix in constructing the compromise, greater weight indicates greater contribution.
2/ "NS norm^2" is the norm squared. The greater the value, the greater will be the first component (i.e., it will bear more information) and may better contribute to the construction of the compromise.
3/ "Cos^2" provides the squared cosines of the angles, which indicates the quality of the representation borne by each matrix in the compromise;
* Outside of confidence interval

An analysis of each of the items determines that items 1, 2 and 3 are the ones that have less correlation with the single vector (one-dimensional factor) constructed from the whole sample without differentiating by country. As a result, we have clear evidence that these items have characteristics differing from the others and that, indeed, are the items that have less factorial load when a single factor is obtained through the PCA.

On the other hand, measurement invariance is assessed according to Brown's criteria (2012) and Cheung & Rensvold (2002) reacting to a progressive approach, first adjusting the model in the 26 countries without any restriction (configuration invariance), then analyzing the invariance in factorial loads (metric invariance), thirdly, an analysis of the invariance in the factorial and intercepted loads (scalar invariance) and, finally, the invariance in the factorial loads, intercepts and variances of the error (residual invariance) is evaluated.

In each model, the established restrictions were added to the parameters of the previous model. To evaluate the invariance, the change in Chi-square was taken into account ($\triangle\chi^2$), with the assumption that the model is invariant if the change is not significant. It was also considered that the change in the coefficient *CFI*($\triangle$*CFI*)not exceed .01 since the χ^2 is very sensitive to sample size and non-normality.

A one-dimensional factorial structure was tested in the total sample (CFI=0.972, AIC=416906.733, BIC=417143.413, RMSEA=0.048, 90% CI 0.046$\leq$*RMSEA*$\leq$*0.05*, SRMR=0.023) and in each of the countries separately. The results (Table 4) indicate that the one-dimensional model presents a good fit to the data only in four countries (Italy, Germany, Costa Rica and Indonesia). Only Germany has excellent 90.0% CI; with indices χ^2/gl, *RMSEA*, *SRMR* and *CFI* that reach fairly good values. Meanwhile, the other countries do not achieve stability in the evaluation criteria, even countries like Greece and France are truly poor, being consistent with previous results in this same study.

Therefore, when considering the previous assertion, it is difficult to sustain the one-dimensional model as a base model for the analysis of invariance. However, in order to test it, the configural invariance was originally estimated.

The indicators of adjustments obtained indicate that the measurement model is not invariant in the different countries, given that not all the adjustment indicators presented values suggested in the literature. Then restrictions were added to the factor loads to evaluate the metric invariance. In this case an acceptable adjustment was not found, leaving further doubts of the unidimensionality, given that there was a significant change between the values *CFI* of the models*($\triangle$CFI>0.01)*, which means that the factorial loads are different in the countries, that is, the items do not have the same weights in the different countries for the latent variable.

Subsequently, the equivalence between the intercepts (scalar invariance) and the indicators show a relative adjustment of this model. However, when analyzing residual invariance by imposing restrictions on the error terms of the items in all 26 countries, the results showed that by restricting the model so that the error terms are equivalent, the adjustment deteriorates significantly when changes occur, significant in the values of *CFI(($\triangle$CFI=0.32)* and in the chi-square values. Consequently, the invariance of the residuals in the groups of the countries analyzed was not corroborated, concluding that the observed data has a lack of invariance.

Regarding the analysis of differential item functioning (DIF), the results point to the evidence that the GSE does not behave in the same way in the different countries; because there are significant differences when analyzing several countries (Table 5). For this analysis, pairs of countries were compared according to the critical value of 0.054 (Rajú, 1999), trying to assess different experiences according to the location of the country in the first and fourth quadrants.

In the case of India and Germany, located at the end of the first quadrant, items 1, 6 and 9 present Differential Item Functioning (DIF). For Spain and Iran, items 2, 3 and 4 have DIF. In the case of Syria and Japan, as well as Spain and Costa Rica in all the items, significant DIF is observed; when comparing Germany and

Table 4
Evaluation of the latent structure and invariance (26 countries) of the GSE scale: Unifactorial Model, 2017

Model	X^2(df)	df	KMO	X^2/df	RMSEA	L090	H190	SRMR	CFI	ΔCFI
All Groups (Ma)	1.654.5	35	0.927	47.271	0.048	0.046	0.050	0.023	0.972	
Italy	41.4	35	0.838	1.182	0.036	0.000	0.073	0.045	0.977	
Indonesia	78.9	35	0.877	2.254	0.048	0.034	0.063	0.034	0.959	
Germany	489.0	35	0.906	13.971	0.043	0.039	0.046	0.023	0.969	
Costa Rica	111.8	35	0.901	3.193	0.048	0.038	0.058	0.031	0.967	
Canada	88.5	35	0.917	2.529	0.065	0.048	0.081	0.033	0.961	
Peru	113.7	35	0.888	3.250	0.048	0.038	0.057	0.032	0.960	
Korea	59.2	35	0.895	1.692	0.069	0.036	0.098	0.046	0.953	
Denmark	61.9	35	0.900	1.769	0.071	0.041	0.099	0.045	0.952	
Russia	118.2	35	0.898	3.377	0.069	0.056	0.083	0.037	0.941	
Great Britain	164.9	35	0.917	4.712	0.091	0.077	0.105	0.042	0.931	
Poland	178.4	35	0.900	5.096	0.077	0.066	0.088	0.041	0.930	
India	77.0	35	0.836	2.200	0.055	0.038	0.072	0.043	0.926	
Portugal	111.4	35	0.842	3.182	0.063	0.050	0.077	0.043	0.912	
USA	549.4	35	0.906	15.697	0.096	0.089	0.103	0.042	0.912	
Iran	247.2	35	0.884	7.063	0.087	0.077	0.097	0.045	0.905	
Hungary	95.9	35	0.878	2.741	0.105	0.08	0.130	0.052	0.903	
Spain	159.4	35	0.880	4.555	0.094	0.08	0.109	0.052	0.901	
Japan	270.5	35	0.918	7.729	0.125	0.111	0.139	0.052	0.900	
Switzerland	164.4	35	0.857	4.698	0.079	0.067	0.091	0.046	0.896	
Hong Kong	401.7	35	0.891	11.476	0.099	0.091	0.108	0.050	0.892	
Belgium	105.0	35	0.860	3.000	0.107	0.084	0.131	0.062	0.886	
Syria	108.5	35	0.833	3.100	0.089	0.070	0.109	0.058	0.872	
Finland	109.8	35	0.872	3.138	0.116	0.092	0.141	0.065	0.863	
Netherlands	472.0	35	0.865	13.487	0.117	0.108	0.127	0.055	0.853	
France	84.2	35	0.825	2.404	0.117	0.085	0.149	0.073	0.848	
Greece	69.9	35	0.782	1.997	0.100	0.065	0.134	0.077	0.829	
ConInv	4,532.1	910		4.980	0.072			0.036	0.930	
MetInv	5,750.5	1135		5.067	0.073			0.062	0.910	0.020**
ScaInv	3,332.8	1135		2.936	0.051			0.056	0.962	-.052***
ParSca	19,269.9	1360		14.169	0.132			0.127	0.642	0.320***
StrInv	25439.6	1610		15.801	0.140			0.136	0.537	0.105***

Note:
Ma = one-dimensional model; χ^2 = chi-square; df = degrees of freedom; RMSEA= root mean square error of approximation; SRMR = standardized root mean residual; Cfi = Comparative Fit index. * <.01. *** p <.001; ConInv = Configural Invariance (Picture); MetInv=Metric Invariance (Loadings); ScaInv = Scalar Invariance Intercepts; ParSca=Partial Scalar Invariance Self1; StrInv = Strict Invariance Error variance

Table 5
NCDIF value of the items that have significant Differential Item Functioning according to critical Rajú value 0.054 (Prob = 0.0000), 2017

Items	Countries pairs					
	India - Germany	Spain - Iran	Syria - Japan	Italy - Korea	Spain - Costa Rica	Switzerland - Germany
Self1	0.11		0.47		2.98	0.09
Self2		0.34	0.92		5.39	0.09
Self3		0.08	1.28		2.02	0.19
Self4		0.10	1.13		5.20	0.34
Self5			1.9		5.24	
Self6	0.63		0.63		5.88	0.32
Self7			1.51		4.15	
Self8			0.96		4.56	0.09
Self9	0.08		0.52		3.25	
Self10			0.13	0.06	5.22	0.17

Switzerland, items 1, 2, 3, 4, 6, 8 and 10 also present significant DIF.

Conclusions

Based on the results obtained, we conclude that a single construct that represents perceived self-efficacy with an explained variability close to 40% shows significant differences that are not observed, reflected in a differential item functioning. Therefore, these results suggest that the GSE (German version) designed in 1981 by Mathias Jerusalem and Ralf Schwarzer with 10 items (Scholz et al., 2002), seems not to represent a one-dimensional and universal factor.

The relative importance of factor loads in different countries when a single factor is formed in a unique way shows that a similar perspective of evaluation of one item or another is lacking, in the different countries studied. The significant differences that are presented in the information curves obtained with IRT show that caution should be used when generalizing the possible use of the scale or evaluating the elimination of items that do not statistically support a single factor, recalibrating the interpretation of the construct.

Bandura recommends that in order to measure self-efficacy we should do so in terms of capacity ("I can"); interestingly, as seen in this research, items 1 and 2 of the scale formulated in this way present differential item functioning. Likewise, we confirmed as Bandura specified that the concept of self-efficacy is multidimensional and that the general aspects are best evaluated with multidimensional scales of self-efficacy and not by multipurpose scales with a few items that try to measure self-efficacy uniformly. In conclusion, Bandura (2012) states that it is better to use multidimensional self-efficacy scales linked to relevant activity domains than through a multipurpose scale with a small set of items.

Finally, it would be very convenient to continue investigating this issue with the application of an alignment method approach to testing for approximate measurement invariance, particularly with a cross-cultural application such as this (Byrne & van de Vijver, 2017).

References

Abdi, H., Williams, L., Valentin, D., & Bennani, M. (2012). STATIS and DISTATIS: Optimum multi-table principal component analysis and three-way metric multidimensional scaling. *Computational Statistics*, *4*(2), 124-167. doi:10.1002/wics.198

Ackerman, T. (1992). A didactic explanation of item bias, item impact, and item validity from a multidimensional perspective. *Journal of Educational Measurement, 29*(1), 67-91. doi: 10.1111/j.1745-3984.1992.tb00368.x

Bandura, A. (1977). Self-efficacy: Toward a unifying theory of behavioral change. *Psychological Review*, *84*(2), 191-215. doi: 10.1037/0033-295X.84.2.191

Bandura, A. (1978). Reflections on self-efficacy. *Advances in Behavior Research and Therapy*, *1*(4), 237-269. doi: 10.1016/0146-6402(78)90012-7

Bandura, A. (1982). Self-efficacy mechanism in human agency. *American Psychologist*, *37*(2), 122-147. doi: 10.1037/0003-066X.37.2.122

Bandura, A. (1986). The explanatory and predictive scope of self-efficacy theory. *Journal of Clinical and Social Psychology*, *4*(3), 359-373. doi: 10.1521/jscp.1986.4.3.359

Bandura, A. (1999). Social cognitive theory: An agentic perspective. *Asian Journal of Social Psychology*, *2*(1), 21-41. doi: 10.1111/1467-839X.00024

Bandura, A. (2012). On the functional properties of perceived self-efficacy revisited. *Journal of Management*, *38*(1), 9-44. doi: 10.1177/0149206311410606

Brown, G. T., Harris, L. R., O'Quin, C., & Lane, K. (2015). Using multi-group confirmatory factor analysis to evaluate cross-cultural research: Identifying and understanding non-invariance. *International Journal of Research and Method in Education*, *40*(1), 66-90. doi: 10.1080/1743727X.2015.1070823

Brown, T., & Moore, M. (2012). Confirmatory factor analysis for applied research. In R. H. Hoyle (Ed.), *Handbook of Structural Equation Modeling* (1st ed., pp. 361-379). New York, NY: The Guilford Press.

Byrne, B. (2001). *Structural Equation Modeling with AMOS Basic Concepts, Applications, and Programming*. New Jork: Psychology Press.

Byrne, B. M., & van de Vijver, F. J. R. (2017). The maximum likelihood alignment approach to testing for approximate measurement invariance: A paradigmatic cross-cultural application. *Psicothema*, *29*(4), 539-551. doi: 10.7334/psicothema2017.178

Carrasco, M., & Del Barrio, M. (2002). Evaluación de la autoeficacia en niños y adolescentes [Assessment of children's and adolescent's self-efficacy]. *Psicothema*, *14*(2), 323-332.

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.

Goldstein, H. (1980). Dimensionality, bias, independence and measurement scale problems in latent trait test score models. British Journal of Mathematical and Statistical Psychology, 33, 234-246. doi: 10.1111/j.2044-8317.1980.tb00610.x

Gómez-Benito, J., Balluerka, N., González, A., Widaman, K. F., & Padilla, J.-L. (2017). Detecting differential item functioning in behavioral indicators across parallel forms. *Psicothema*, *29*(1), 91-95. doi: 10.7334/psicothema2015.112

Hambleton, R. K. (1996). Guidelines for adapting educational and psychological tests: A progress report. *European Journal of Psychological Assessment, 10*(3), 229-244. doi: 10.1027/1015-5759.11.3.147

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. doi: 10.1080/10705519909540118

Jolliffe, I. T. (2002). Rotation and interpretation of principal components. In Springer (Ed.), *Principal Component Analysis* (2nd ed., Vol. 30, p. 487). Springer, New York, NY.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. *Applied Psychological Measurement* (Vol. 5). New York: Lawrence Erlbaum Associates, Hillsdale. N.J. doi: 10.1177/014662168100500412

Lorezo-Seva, U., & Ferrando, P. J. (2011). *Factor 8* (Release 10.8.03) [Manual of the program]. Tarragona, Spain: Universitar Rovira i Virgili.

Muñiz, J. (2010). Test Theories: Classical Theory and Item Response Theory. *Papeles del Psicólogo, 31*(1), 57-66.

Ning-Min, S., & Jing, L. (2015). A literature survey on high-dimensional sparse principal component analysis. *International Journal of Database Theory and Application*, *8*(6), 57-74. doi: 10.14257/ijdta.2015.8.6.06

Rajú, N. S. (1999). DFITP5: A Fortran program for calculating dichotomous DIF/DTF [Computer software]. Chicago: Illinois Institute of Technology.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph Supplement.*, *34*(Supplement 1), 1-97.

Scholz, U., Gutiérrez, B., Sud, S., & Schwarzer, R. (2002). Is general self-efficacy a universal construct? *European Journal of Psychological Assessment*, *18*(3), 242-251. doi: 10.1027//1015-5759.18.3.242

Schwarzer, R., & Jerusalem, M. (1995). Measures in Health Psychology: A user's portfolio. Causal and control beliefs. *Causal and Control Beliefs, 1*, 35-37.

Sireci, S., & Padilla, J. L. (2014). Validating assessments: Introduction to the Special Section. *Psicothema, 26*(1), 97-99. doi: 10.7334/psicothema2013.255

Suárez-Álvarez, J., Pedrosa, I., Lozano, L. M., García-Cueto, E., Cuesta, M., & Muñiz, J. (2018). Using reversed items in likert scales: A questionable practice. *Psicothema*, *30*(2), 149-158. doi: 10.7334/psicothema2018.33

Spearman, C. (1904). The proof and measurement of association between two things. *The American Journal of Psychology*, *15*(3), 72-101.

Thissen, D., Chen, W. H., & Bock, R. D. (2003). MULTILOG 7: Multiple categorical item analysis and test scoring using item response theory (Version 7.03) [Computer software]. Chicago: Scientific Software International.

Thurstone, L. L. (1931). Multiple factor analysis of variance. *Psychological Review*, *38*, 406-427.

Zou, H., Hastie, T., & Tibshirani, R. (2006). Sparse principal component analysis. *Journal of Computational and Graphical Statistics*, *15*(2), 265-286. doi: 10.1198/106186006X113430

Zou, H., & Hastie, T. (2015). Elasticnet: Elastic-Net for Sparse Estimation and Sparse PCA (Version 1.1) [Computer software]. Minnesota: School of Statistics.