Revista de Psicopatología y Psicología Clínica, 23, **211-220, 2018**
doi: 10.5944/rppc.vol.23.num.3.2018.21160
http://www.aepcp.net
http://revistas.uned.es/index.php/rppc
Rev. Psicopatol. Psicol. Clin.
ISSN 1136-5420

Rasch model analysis of the Brief Version of the Social Phobia and Anxiety Inventory (SPAI- B) in Argentinean and Spanish samples

Valeria E. Morán[1], Marcos Cupani[1], Ana E. Azpilicueta[1], José A. Piqueras[2], and Luis J. Garcia-Lopez[3]

[1]*Universidad Nacional de Córdoba, Córdoba, Argentina*
[2] *Universidad Miguel Hernández de Elche, Alicante, Spain*
[3] *Universidad de Jaén, Jaén, Spain*

Abstract: A valid and reliable assessment of social anxiety is a subject of practical relevance in the field of clinical research and diagnosis. In the present study, psychometric properties of the Spanish version of the Social Phobia and Anxiety Inventory (SPAI-B) using the item response theory (IRT) were assessed. Although the inventory is widely used in several countries and cultural contexts, validation studies of the scale had not been performed yet from the Rasch model perspective in Argentinean and Spanish populations. The results indicate that the tool allows a unidimensional assessment of social anxiety with a pertinent categorization of the response scale. Furthermore, the data show that the level of severity of the items is adequate and allows the detection of clinical population. It was reported on the items with a differential functioning between gender and nationality, and the clinical implications of such differences were discussed.

Keywords: Social anxiety; SPAI-B; item response theory; Rasch model.

Análisis mediante el modelo de Rasch de la Versión Breve del Inventario de Ansiedad y Fobia Social (SPAI-B) en muestras argentinas y españolas.

Resumen: La evaluación válida y fiable de la ansiedad social es un aspecto de suma importancia práctica en contextos de investigación y diagnóstico clínico. Por ello, en el presente estudio se evaluaron las propiedades psicométricas de la versión breve para hispanohablantes del Social Phobia and Anxiety Inventory (SPAI-B) a partir de la teoría de respuesta al ítem. Si bien la escala es ampliamente utilizada en diversos países, contextos sociales y culturales, hasta ahora no se habían llevado a cabo estudios de validación de la escala desde el modelo de Rasch en población española y argentina. Los resultados evidencian que el instrumento permite una evaluación unidimensional de la ansiedad social con una categorización pertinente de la escala de respuesta. Asimismo, los datos muestran que el nivel de severidad de los ítems es adecuado y permite la detección de población clínica. Por último, se informa sobre los ítems que tienen funcionamiento diferencial entre sexo y nacionalidad, y se discuten las implicaciones clínicas de estas diferencias culturales.

Palabras clave: Ansiedad social; SPAI-B; teoría de respuesta al ítem; modelo de Rasch.

Introduction

Beginning a university career happens simultaneously with a series of essential life events, such as leaving the family home and looking for a job and a partner. This conjunction of events, which dares young people to face several emotional and interpersonal demands and exposes them to new environments, can be threatening if they do not have the necessary resources to deal with it (Morán, 2018). Within this context, students are continuously evaluated not only by university teachers and authorities but also by members of the group of students they belong to and to whom they relate daily (Velásquez et al., 2008). From this perspective, it is undeniable the importance of the social relationships

Received: September 3, 2018; accepted: November 27, 2018
Corresponding author: Valeria E. Morán, Facultad de Psicología, Universidad Nacional de Córdoba, Enfermera Gordillo s/n, Ciudad Universitaria, 5000 Córdoba, Argentina. Email: moranvaleria@gmail.com

along the attendance of a university career, in line with the findings of Rubin, Evans and Wilkinson (2016), who stated that both social contact and subjective social status predict positively the mental health and wellbeing of university students.

The Social Anxiety Disorder (SAD) is one of the pathologies specifically related to interpersonal relationships that might affect university students. The SAD is a pathology characterized by excessive and persistent fear to negative evaluation in social interaction situations, accompanied by a simultaneous fear of feeling embarrassed or humiliated, which might lead to social rejection (American Psychiatric Association, 2013). Consequently, this suffering is associated to a significant clinical discomfort and to a meaningful interference in people's daily life in the social, academic and working environments.

Knappe, Sasagawa and Creswell (2015), in their review on epidemiological studies, stated that the SAD is the second most usual disorder out of all the ones included in the DSM-5, with prevalence rates around 13% of the population. However, the percentage of young adults who report social fears but are not diagnosed with SAD is notably higher and ranges between 27% and 47% (Essau, Condrat, & Petermann, 2000). The increase of both the epidemiological studies and the research applied to social anxiety has been accompanied by the development of tools designed to assess the disorder (Alcántara-Jiménez & Garcia-Lopez, 2017). One of the most widely used tools in the Spanish-speaking population is the brief version of the Social Phobia and Anxiety Inventory, SPAI-B (Garcia-Lopez, Hidalgo, Beidel, Olivares, & Turner, 2008), developed from the Social Phobia and Anxiety Inventory, SPAI (Turner, Beidel, Dancu, & Stanley, 1989), which was designed to assess social anxiety considering its somatic, behavioral, and cognitive aspects. The brief version not only presents a reduction of items and administration time but also modifies the response range by adopting a Likert scale of five points (two points less than the original version with seven options). One of the advantages of this version is that it avoids heterocentric language, allowing a social anxiety measurement more adequate for social minorities such as the collective LGTBi, among others.

The version adapted to the Spanish population presents adequate psychometric properties, such as internal structure validity and temporal stability (Garcia-Lopez et al., 2008), sensitivity to therapeutic change (Garcia-Lopez, Díaz-Castela, Muela-Martínez, & Espinosa-Fernández, 2014), and good discriminant power between population control and SAD, including a cut-off point differentiated for the social performance specifier described in the DSM-5 (Garcia-Lopez, Beidel, Muela-Martinez, & Espinosa-Fernandez, 2016; Garcia-Lopez, Saez-Castillo, Beidel, & La Greca, 2015). Besides, the scale also shows validity evidence for groups compared in terms of gender, being the found differences in favor of women with small size effects (Garcia-Lopez et al., 2008, 2015; Piqueras Espinosa-Fernández, Garcia-Lopez, and Beidel, 2012), which is consistent with other studies on SAD that support the absence of necessity of different cut-off points in terms of gender (for a review, see Garcia-Lopez, Salvador, and De Los Reyes, 2015).

The SPAI-B has recently been adapted to be used in the Argentinean university population (Morán, Azpilicueta, Cupani, & Garcia-Lopez, 2018), presenting adequate reliability levels (ω= .89) and a unifactorial structure similar to the original scale validated in Spanish university students. Besides, the Argentinean version presented significant correlations with the punctuations of the Test of Social Anxiety (TAS-U for its acronym in Spanish; Morán, Olaz, Pérez & Del Prette, 2018), showing its convergent validity. The psychometric properties assessed in the different versions of the scale have presented satisfactory results. However, the performed analyses are based on the classical test theory (CTT), which entails a problem of double invariance that should be considered, among others. On the one hand, the properties of the test and its items are relative to the sample of individuals studied; on the other hand, the person measurements are relative to the used tool, and vice versa (Cupani & Cortez, 2016).

The item response theory (IRT) presents a series of analysis models to solve some of the disadvantages of the CTT. It ensures the invariance of the measurements allowing the independence in the analyses between individuals and tool (Engelhard Jr., 2013). On the other hand, the measurement precision is estimated for each ability level in the variable, emphasizing the item analysis and the level of a person's ability by emitting a response to them; this is the reason why the measurement error is calculated for each item and for each person (Cupani & Cortez, 2016). Within the different analysis models presented by the IRT, the Rasch model allows to assess the psychometric properties of the tools according to the properties of each item of the test (Messick, 1994). Specifically, it provides information about the ability of a person to respond adequately to the item instead of reporting about the expected number of responses. According to the Rasch model, the score for each subject is the result of the interaction between the person's ability

and the difficulty of the item (Linacre, 2002). Applying this model allows the data matrix to provide sufficient statistics for the calculation of the parameters of the persons (β) and the items (δ) (Andersen, 1977), as long as they are single-variable locations, and can be expressed in the same scale unit, allowing to establish objective comparisons (Lozano, 2005).

The nature of the SPAI-B variable requires the use of rating scales (e.g. Likert-type) with preference to any other method. One model which fits the characteristics of this variable is the Rating Scale Model (RSM). This is a polytomous model derived from Rasch dichotomous model. The RSM requires that the response categories be gradual, determined, exhaustive and excluding (Andrich, 1978). It also allows to estimate the probability of a person response to a certain category of an item, considering the difference between the level of ability of the person in the variable being measured and the capacity of the items for measure the variable. According to the model, the probability of a person's response to an item is expressed by the following equation:

$$P_{nix} = \frac{\sum_{j=1}^{k}(\beta_n - (\delta_i + \tau_j)}{1 + \sum_{k=1}^{m} exp \sum_{j=1}^{k}(\beta_n - (\delta_i + \tau_j))} \qquad x = 1,2,3 \ldots m$$

where βn is the location or scale value of the person n; δi is the location or scale value of item i; and τj is the threshold parameter between categories and represents the locations of step m relative to the scale value of the item. The only difference between items is due to their different location (δi) in the one-dimensional continuum of the variable that is measured (Masters, 1988). The values of τj must be kept constant across all the items that form the scale, and it is assumed that they depend only on the proposed response categories (Wright & Masters, 1982). The application of this model implies the data empirically obtained fits the prediction of the model. That is why before interpreting the results it is necessary to check the fit of the data to the model.

Following these trends and with the aim of optimizing the tool, this work aims to analyze the SPAI-B properties from the IRT. Due to these advantages, the IRT and particularly the Rash model counteract the CTT constraints and are currently the most widely used tools in the construction of scales and tests, especially in the health area (Embretson & Reise, 2000; Hagquist & Andrich, 2017; Hoertel et al., 2016; Rivollier et. al., 2015). Gómez-Benito, Sireci, Padilla, Hidalgo & Benítez (2018), which indicates that this model analysis constitutes a method to obtain evidence of all the validity sources: internal structure, content, response process, relationships to other variables, and consequences of the assessment. Specifically, the Rasch model also informs about the influence of certain variables at the time of emitting an expected or correct response by analyzing the differential item functioning (DIF) (De Ayala, 2009). In order to assess these differences, we verify whether there is a DIF between the Argentinean and the Spanish samples, and the male and female participants. In addition, the effectiveness of the response categories, the unidimensionality, the fit of the model and the separation and reliability of both items and persons, were analyzed.

Method

Participants

A sample of 635 students (77 % women, 23 % men, age range: 18–49 years, M_{age} = 22.19 years, SD = 3.33) were recruited in universities in Spain (61 %) and Argentina (39 %). We used the non-probability sampling method of accidental-type. The gender ratio presented in this study is consistent with the data found for both countries.

Instruments

Brief form of the Social Phobia and Anxiety Inventory, SPAI-B (Garcia-Lopez et al., 2008): It is a questionnaire to assess social phobia symptoms in adolescents, consisting of 16 items with a Liker-type response scale of five points. Participants responded about the frequency of occurrence in the situation described in the item (1 = *Never*, 5 = *Always*). The exploratory and confirmatory factorial analysis evidenced a unifactorial structure with an excellent internal consistency in both, the Spanish samples (α = .89) (Piqueras et al., 2012) and the Argentinean samples (α =.85) (Morán, Azpilicueta et al., 2018). In this study we obtained an α = .89.

Procedure

Participants, after being informed about the aim of the study, signed a consent form that stated the aim, the voluntary participation, and the confidential nature of the data. Psychology students, trained for the purpose, collected the data at the universities.

Data Analysis

We worked with the software Winsteps (Linacre, 2016). The Rating Scale Model (RSM) was applied for

polytomous items. The three main assumptions of this model are unidimensionality, that is, all the items must measure the same construct; the local independence of the items and person, which implies that the response of an individual to any of the test items will not be affected by their response in another item, which also occurs with the items; and that all the items have the same power of discrimination (Andrich, 1978). The item calibration program consisted of the following steps:

Effectiveness of the rating categories: The first step was to examine the point-biserial correlations to ensure that all the elements were oriented in the same direction, that is, corroborate that all correlations were positive (Linacre, 2002). The second step was to evaluate the accuracy of the classification scale, based on certain criteria. First, in order to estimate threshold values in a stable manner, at least 10 observations must be presented for each rating category. Second, the average measures for each category should increase uniformly as the variable increases. Third, the thresholds, that is, the estimated difficulties in choosing one response category over another, must also increase uniformly across the rating scale, so as not to be considered as disordered. Fourth, each step between categories must define a different range in the variable, which is observed in the magnitudes of the distances between the category thresholds. Linacre (2002) suggests that the thresholds should increase at least 0.81 logits for a scale of 5 points, in order to show the distinction between the categories, but no more than 5 logits, to avoid large gaps in the variable. Finally, the statistical fit provides another criterion for evaluating the quality of a classification scale. However, when an Outfit value greater than 2 is obtained, it indicates more misinformation than information, so the category introduces interference in the measurement process. These criteria are generally used in combination, to detect any disordered categorization (problematic categories), that are feasible to collapse, and to determine the optimal categorization.

Unidimensionality: The unidimensionality of the scale was evaluated by using Rasch Principal Components Analysis of Residuals (PCAR). We considered that the unidimensionality assumption is achieved if the measurement model explains approximately the 50 % of the variance, if the first contrast (a secondary dimension) has its eigenvalue smaller than 3 (a force of three items) and explains less than 5 % of the unexplained variance (Linacre, 2016).

Rasch model fit: First, the global fit of the data was analyzed in order to check if the data matrix fits the predictions of the model. Second, the items fit was analyzed by studying each one independently. Third, the persons fit was analyzed in order to identify the participants who responded unexpectedly and not adequate to the theoretical formulation. We used the statistical fit indexes Infit (internal fit) and Outfit (external fit). The Infit index is calculated from the unstandardized quadratic means, which allows identifying unexpected behaviors that affect the items that, in the measurement continuum, are close to the level of trait that a person possesses. The Outfit index is the weighted root mean of residuals resulting from persons and items; it allows to evaluate the unexpected behavior of the items that have a difficulty far from the level of latent trait that each person presents (Bond & Fox, 2015). Values provided by the Rasch Model are expressed in logit scale, which is a logistic transformation of the observed scores, with a mean of 0 and standard deviation of 1. When the fit is adequate, that is, the observed data match with those proposed by the model, the values of Infit and Outfit are close to 1; otherwise, values far from 1 will be obtained. This means that an Infit value of 1 indicates that 100% of the variance of the empirical data is explained by the model, while a value greater than 1.3 indicates that there is more variance of the expected (30% of the variance) that cannot be explained by the model. Following the criteria proposed by Wang and Chen (2005) and Wright and Linacre (1994), the region to consider an acceptable fit oscillates between 0.6 and 1.4 logits for polytomous items.

Separation and reliability. The item separation index indicates the distance between the levels of difficulty or trait, which should be sufficient to identify the meaning of the latent variable (Wright & Stone, 2003). The person's separation index indicates the aptitude of the instrument to discriminate people in the measured variable. A useful set of items must define at least three strata of people (e.g., high, moderate and low levels).

An adequate level of separation should be greater than 2 (Bond & Fox, 2015), associated to a reliability around 0.80 (Bond & Fox, 2003). To assess the position of items and persons in the continuous, we analyzed the map of items and persons simultaneously.

Differential item functioning (DIF): We studied the DIF in terms of gender and nationality of the participants. An item is considered to have DIF when the probability of a correct response does not depend solely on the level of the person in the ability/trait intentionally measured by the test (Bond & Fox, 2015). Item severity measures δi were computed for each class of subjects (e.g., male vs. female/ Spanish vs Argentinean). A two-sided t test was then performed to pairwise compare item severity measures between subject classes. Significance level was set by Bonferroni´s correction for multiple comparisons.

To attribute the differences to the grouping variables, we considered that the DIF contrast (e.g., DIF measure for subject class 1 minus subject class 2) should be ≥ 0.5 logits (Linacre, 2016).

To evaluate the impact of the independent variables, an ANOVA was performed with the logit scores of the level of severity of each person and the effect size (partial eta squared) was calculated for the significant differences ($p \leq .001$), considering the proposed criteria by Cohen (1988), being a small effect when $\eta^2 = .01$, medium $\eta^2 = .061$ and large for $\eta^2 = .14$.

Results

Effectiveness of the rating categories

The point-biserial correlation for the 16 items was positive, indicating that the items were written in the same direction and measured the same construct. As regards the classification structure, all the criteria proposed by Linacre (2002) were fulfilled: frequencies were high for all the categories, the average increase of the measurement was near to .80 along the categories, the Outfit values were close to 1 for all the categories, and the thresholds also increased uniformly, which would indicate that each category is the most likely for a specific range in the construct continuum. Besides, the distances between the consecutive thresholds had enough magnitude to describe different ranges in the variable (see Table 1).

Unidimensionality

We examined the unidimensionality of the tool through the analysis of the PCAR. The results indicated that the assumption of unidimensionality was accomplished because the Rasch dimension explained 66.20 % of the of the data. The first contrast (the highest secondary dimension) had an eigenvalue of 2.10 and represented the 4.30 % of the unexplained variance.

Rasch model fit

The Infit and Outfit values were within the expected ranges, indicating an adequate fit of the 16 items (see Table 2). The item severity measurements (δi) ranged between $-1.73 \leq \delta i \leq 1.18$, with an average of 0.00 (*SD* = 0.80). The Infit values of the items ranged between 0.73 y 1.42 and the Outfit ones ranged between 0.72 and 1.36, with an average of 1.02 (*SD* = 0.23) and 1.01 (*SD* = 0.23), respectively. As regards the fit for persons, we found that 55 % of the response patterns fit to the model (I*nfit* ≥ 0.60 and *Outfit* ≤ 1.40). However, it is possible to indicate that 93% of the participants presented an acceptable fit if we consider that only 43 individuals obtained severe misfit ($\theta > 2$) according to Linacre (2016) criteria. The severity levels varied between $-4.92 \leq \theta \leq 2.68$, with an average of -1.05 (*SD* = 1.05). Figure 1 shows the map of Wright with the adjacent distribution of persons and items. The bottom of the map (negative values) corresponds to the lower level of severity of the persons and items, while moving up in the map (positive values) corresponds to higher level of severity. In the figure it can be seen that most of the items are located in a centered position with respect to the students evaluated, and that the items, in general lines, achieve an adequate distribution through the continuum. The data also indicate that the test evaluates high levels of social anxiety severity (average of $\delta = 0.00$) for the sample of students analyzed (average of $\theta = -1.05$), therefore, it could be affirmed that those who score high in this test have high levels of the construct.

Separation and reliability

The item separation index (15.31), the item reliability index (1.00), the values of separation indexes of persons (2.84), and the reliability of persons (0.89) were satisfactory. These results indicated that the size of the used sample was enough to establish the hierarchy of item difficulty of the tool (Linacre, 2016).

Table 1. Comparison of the classification categories

Category	Observed counting	Average measurement	Outfit mean square	Calibration
1	2865	-2.22	0.99	None
2	3496	-1.19	0.86	-1.87
3	2294	-0.39	0.94	-0.40
4	1016	0.33	1.05	0.76
5	457	0.95	1.39	1.51

Table 2. Fit of items, measurements and persons

	Measurement	Standard error	MNSQ INFIT	MNSQ OUTFIT	*Rpb*[1]
Summary of persons					
M	-1.05	0.34	1.01	1.01	
SD	1.05	0.08	0.58	0.61	
Maximum	2.68	1.03	3.75	4.47	
Mínimum	-4.92	0.28	0.20	0.19	
Summary of Items					
M	0.00	0.05	1.02	1.01	
SD	0.80	0.01	0.23	0.23	
Maximum	1.18	0.06	1.42	1.36	
Minimum	-1.73	0.05	0.73	0.72	
Item1	-.82	.05	.77	.77	.62
Item2	-.68	.05	.75	.74	.63
Item3	-1.73	.05	1.26	1.30	.45
Item4	1.18	.06	1.17	1.08	.49
Item5	-.31	.05	1.02	.99	.61
Item6	-.33	.05	.94	.90	.57
Item7	.13	.05	1.17	1.16	.50
Item8	.10	.05	.77	.75	.65
Item9	.90	.06	1.42	1.33	.45
Item10	-.03	.05	.98	.99	.57
Item11	-1.18	.05	1.34	1.35	.45
Item12	.84	.06	1.35	1.36	.45
Item13	-.10	.05	.73	.72	.65
Item14	.30	.05	1.05	1.01	.59
Item15	.72	.06	.73	.78	.63
Item16	1.01	.06	.86	.86	.57

Note: [1] *Rpb* = point-biserial correlation.
MNSQ = mean square; INFIT = internal fit;
OUTFIT = external fit.

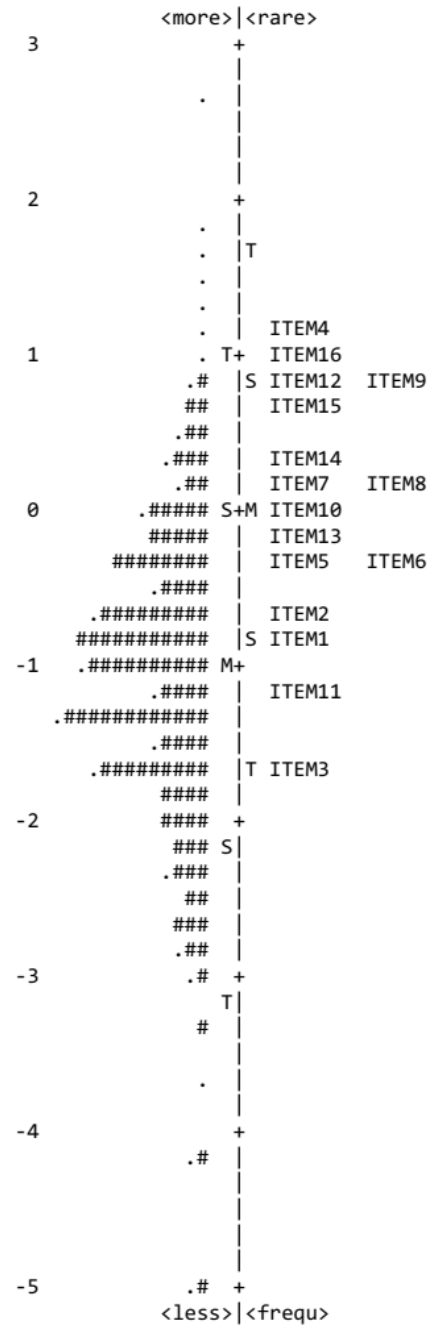

Figure 1. Person-item map. The left-hand column shows the person location along the continuum considering ability level. The symbol # represents a group of five persons and symbol … represents one person. This distribution usually takes the shape of a normal curve. M marks the media of persons and items. S is a SD far from the media and T is two SDs far from the media.

¡Differential item functioning and impact of external variables

Considering participant genders, the data revealed there are not differences in the items. On the other hand, the analysis of participant nationalities showed significant differences in items 3 ($p < .0031$) with a contrast statistic of 0.52, and a median effect size of .65 (DIF/SD measure). More specifically, the item 3 severity (DIF media) for the Argentinean sample was -1.42 logits, whereas for the Spanish one was -1.95 (Table 3).

Table 3. DIF contrast by gender and nationality

	Gender	Nationality	*d*
Item1	0.01	0.19	
Item2	0.16	0.25	
Item3	0.25	0.52**	0.65
Item4	0.22	0.25	
Item5	0.15	0.27	
Item6	0.30	0.02	
Item7	0.23	0.26	
Item8	0.02	0.38	
Item9	0.23	0.54	
Item10	0.23	0.47	
Item11	0.20	0.19	
Item12	0.48	0.14	
Item13	0.04	0.01	
Item14	0.05	0.13	
Item15	0.20	0.28	
Item16	0.28	0.37	

Note: DIF = Differential item functioning.
d = Cohen's effect size.
$^{**}p \leq .001$.

The analysis of the impact of independent variables on the levels of severity (θ) of persons in the latent variable resulted in significant differences according to gender ($p \leq .001$) and to nationality ($p \leq .01$) but in all cases the effect size was small or null (Table 4). Therefore, the interaction between the two independent variables did not have an impact on the levels of social anxiety.

Discussion

Social anxiety is currently a problem with high impact on social functioning and in academic and working performances of university students (Garcia-Lopez, Diez-Bedmar, and Moreno-Almansa, 2013). Therefore, it is relevant to deepen the knowledge of this phenomenon in that population, for which it is essential to count with valid and precise assessment tools. The aim of this work was to analyze the SPAI-B properties from the IRT considering the advantages provided by this analysis information.

Table 4. Social anxiety severity ANOVA by gender and nationality

		M (SD)[a]	*F*	Partial η^2
Nationality	Argentina	-0.94 (1.02)	6.64**	0.01
	Spain	-1.16 (1.15)		
Gender	Male	-1.28 (1.19)	10.34***	0.02
	Female	-1.01 (1.07)		
Nationality*Gender		—	0.01	—

Note: [a] Statistics in logit scores (θ)
$^{**}p \leq .01$; $^{***}p \leq .001$.

From the obtained results, we can conclude that the scale presents adequate psychometric properties. In the first study, we analyzed the effectiveness of the categories of response classification. The results show that the distances between the consecutive thresholds were wide enough to describe different ranges in the measured variable, which demonstrates that the categories are gradual, exhaustive, and exclusive, representing quantities or increments in the measured variables along the items (Andrich, 1978). Then, we corroborated the unidimensionality of the scale, which had been already evidenced in previous psychometric studies via exploratory and confirmatory factorial analyses (Garcia-Lopez et al., 2008; Morán, Azpilicueta et al., 2018; Piqueras et al., 2012).

On the other hand, we observed that 93 % of the participants responded the test items coherently. The person-item distribution map allowed to compare the difficulty indexes and the ability levels of the respondents. The map shows that the distribution of item difficulties was not completely aligned to the distribution of the person ability levels. Specifically, the lowest person severity levels, that is, those with lower social anxiety, were not represented by the items forming the scale. Considering that this is a diagnostic assessment tool for a clinical construct, the aim is that it can represent and detect the highest levels of the variable because they would evidence the presence of the pathology, and not the lowest ones. Hence, the results show that the difficulty levels of the items exceed the severity or trait levels of the general population, which allows to conclude that the

tool is adequate to identify clinical population. However, we observed that the lower levels of the latent variable were not represented enough in the item continuum. This fact highlights the need of incorporating items to achieve a better coverage of the test.

The analysis of the items shows an optimal global fit for all the reagents; the estimation error for all of them was relatively low (0.05 logits) accounting for the precision of the scale. Moreover, the reliability indexes for both the items and the persons would be predictably reproducible (Andrich, 2002). On the other hand, the separation indexes of the items showed that they present enough distance between the difficulty levels of the latent variable (Wright and Stone, 2003). The separation of the person indexes evidenced the aptitude of the tool to discriminate the persons in the measured variable.

As regards the DIF, we analyzed differences related to nationality and gender and found that just the item 3 (I feel nervous when I have to speak in public) presented significant severity differences between Argentinean and Spanish may be as a reflection of cultural differences. However, it would be pertinent to replicate these analyzes using complex models where fixed parameters are considered for the invariant items and free parameters for item 3. If the validation study confirms DIF for the item, it is important to take into account that when interpreting the sum score across all items of the scale, DIF on a single item probably not cause substantial test-wise bias (Tennant and Pallant, 2007).

Although the present results are satisfactory, it is worthy to mention that the drawback of the performed studies is that the participants were selected through non-probabilistic sample procedures, which might affect the estimation of the item parameters However, efforts were made to include public and private universities, different disciplines and students of different levels of education, in order to reduce this bias.

We showed in this work that the 16 items can be used to assess social anxiety precisely, because they measure the proposed construct and whether the persons have the abilities analyzed by the tool. Moreover, considering that the item distribution located towards the upper end of the person abilities, this tool is useful as a diagnostic screening instrument for SAD. However, in further research we suggest to assess the item distribution in clinical populations to determine whether they are properly represented in all the possible clinical levels in the social anxiety continuum.

As a final conclusion, the results of this work contribute to reinforce the relevance of implementing analyses from the Rasch model in the construction and/or adaptation of tests because that is the most rigorous psychometric model to analyze, assess, and validate the measurement tools.

Funding: This research received especial grant from Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET)

Conflicts of interest

The authors have no conflicts of interest to declare.

References

Alcántara-Jiménez, M. M., y García-López, L. J. (2017). Revisión de los procedimientos observacionales y cognitivos para la evaluación de la ansiedad social [A revision of observational and cognitive assessment measures for a population with social anxiety]. *Revista de Psicopatología y Psicología Clínica*, *22*(3), 243-260. doi: 10.5944/rppc.vol.22.num.3.2017.19187

American Psychiatric Association. (2013). *Diagnostic and Statistical Manual of Mental Disorders* (5th Edition). Arlington, VA: American Psychiatric Association.

Andersen, E.B. (1977). Sufficient statistics and latent trait models. *Psychometrika*, 42(1), 69-81. doi:10.1007/bf02293746

Andrich, D. (1978). Application of a psychometric rating model to ordered categories which are scored with successive integers. *Applied Psychological Measurement, 2*(4), 581-594. doi: 10.1177/014662167800200413

Andrich, D. (2002). Implications and applications of modern test theory in the context of outcomes based education. *Studies in Educational Evaluation*, *28*(2), 103-121. doi:10.1016/S0191-491X (02)00015-9

Bond, T. G. & Fox, C. M. (2003). Applying the Rasch model: Fundamental measurement in the human sciences. *Journal of Educational Measurement*, *40*(2), 185-187. doi:10.1111/j.1745-3984.2003.tb01103.x

Bond, T. & Fox, C. M. (2015). *Applying the Rasch model: Fundamental measurement in the human sciences*. New York: Routledge.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Cupani, M. & Cortez, F. D. (2016). Análisis psicométricos del subtest de razonamiento numérico utilizando el modelo de Rasch [Psychometric analysis of the subtest of numerical reasoning using the Rasch model]. *Revista de Psicología, 25*(2), 1-16. doi:10.5354/0719-0581.2016.44558

De Ayala, R. J. (2009). *The theory and practice of item response theory*. New York, New York: The Guilford Press.

Embretson, S. E., & Reise, S. P. (2013). *Item response theory*. Mawah, NJ: Psychology Press.

Engelhard Jr, G. (2013). *Invariant measurement: Using Rasch models in the social, behavioral, and health sciences*. New York: Routledge.

Essau, C. A. Condrat, J., & Petermann F. (2000). Frequency, comorbidity, and Psychosocial impairment of anxiety disorder in German adolescents. *Journal of Anxiety Disorders, 14*(3), 263-278. doi:10.1016/s0887-6185(99)00039-0

García-Lopez, L.-J., Beidel, D., Muela-Martinez, J. A., & Espinosa-Fernandez, L. (2016). Optimal cut-off score of Social Phobia and Anxiety Inventory-Brief Form. *European Journal of Psychological Assessment*, *34*(4), 278-282. doi:10.1027/1015-5759/a000324

García-Lopez, L. J., Díaz-Castela, M., Muela-Martinez, J. A., & Espinosa-Fernández, L. (2014). Can parent training for parents with high levels of expressed emotion have a positive effect on their child's social anxiety improvement? *Journal of Anxiety Disorders, 28*(8), 812-822.doi: 10.1016/j.janxdis.2014.09.001

García-López, L. J., Díez-Bedmar, M. B., & Almansa-Moreno, J. M. (2013). From Being a Trainee to Being a Trainer: Helping Peers Improve their Public Speaking Skills. *Revista de Psicodidáctica, 18*(2), 331-342.

García-Lopez, L. J., Hidalgo, M. D., Beidel, D. C., Olivares, J., & Turner, S. (2008). Brief form of the Social Phobia and Anxiety Inventory (SPAI-B) for adolescents. *European Journal of Psychological Assessment, 24*(3), 150-156. doi: 10.1027/1015-5759.24.3.150

García-Lopez, L. J., Sáez-Castillo, A. J., Beidel, D., & La Greca, A. M. (2015). Brief measures to screen for social anxiety in adolescents. *Journal of Developmental & Behavioral Pediatrics, 36*(8), 562-568. doi: 10.1097/dbp.0000000000000213

García-Lopez, L. J., Salvador, M.C., & De Los Reyes, A. (2015). Assessment of social anxiety in adolescents. In K. Ranta, A.M. La Greca, L.J. García-Lopez, & Marttunen, M. (Eds), *Social anxiety and phobia in adolescents* (pp. 121-150). Springer International Publishing.

Gómez-Benito, J., Sireci, S., Padilla, J. L., Hidalgo, M. D., & Benítez, I. (2018). Differential item functioning: Beyond validity evidence based on internal structure. *Psicothema, 30*(1), 104-109.

Hagquist, C. & Andrich, D. (2017). Recent advances in analysis of differential item functioning in health research using the Rasch model. *Health and Quality of Life Outcomes*, *15*(1), 181. doi: 10.1186/s12955-017-0755-0

Hoertel, N., Blanco, C., Peyre, H., Wall, M. M., McMahon, K., Gorwood, P., ... & Limosin, F. (2016). Differences in symptom expression between unipolar and bipolar spectrum depression: Results from a nationally representative sample using item response theory (IRT). *Journal of Disorders*, *204*(1), 24-31. doi: 10.1016/j.jad.2016.06.042

Knappe, S., Sasagawa, S., & Creswell, C. (2015). Developmental epidemiology of social anxiety and social phobia in adolescents. In K. Ranta, A. M. La Greca, L. J. García-Lopez, & M. Marttunen (Eds.), *Social Anxiety and Phobia in Adolescents* (pp. 39-70). New York: Springer International Publishing.

Linacre, J. M. (2002). Optimizing rating scale category effectiveness. *Journal of Applied Measurement*, *3*(1), 85-106. doi:10.1.1.424.2811

Linacre, J. M. (2016). Winsteps® (Version 3.92.0) [Computer Software]. Beaverton, Oregon: Winsteps.com. Retrieve from http://www.winsteps.com/

Lozano, O. (2005). *Construcción de un test para medir la calidad de vida relacionada con la salud en drogodependientes. Aplicación de un modelo politómico de la teoría de respuesta al ítem*. [Construction of a test to measure the quality of life related to health in drug addicts. Application of a polithomic model of the item response theory] (Doctoral dissertation). Universidad de Granada, Granada, España. Retrieve from http://hera.ugr.es/tesisugr/15480173.pdf

Masters, G. N. (1988). Measurement models for ordered response categories. In R. Langeheine, & J. Rost, (Eds.). *Latent trait and latent class models* (pp. 11-29). Plenum Press, New York.

Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher, 23*(2), 13-23. doi:10.3102/0013189X023002013

Morán, V. E., Olaz, F. O., Pérez, E. R., & Del Prette, Z. A. P. (2018). Desarrollo y validación del Test de Ansiedad Social para estudiantes universitarios (TAS-U) [Development and validation of the Social Anxiety Test for university students (SAT-U)]. Manuscript submitted for publication.

Morán, V. E., Azpilicueta, A. E., Cupani, M., & García-Lopez, L. J. (2018). *Análisis psicométrico del Inventario de Fobia y Ansiedad Social- Forma Breve (SPAI- B) en muestras argentinas* [Psychometric analysis of the Social Phobia and Anxiety Inventory - Brief Form (SPAI-B) in Argentinean samples]. Manuscript submitted for publication.

Morán, V. E., Olaz, F. O., Pérez, E. R., & Del Prette, Z. A. P. (2018). Emotional-evolutional model of social anxiety in university students. *International journal of psychology and psychological therapy, 18*(3), 315-330.

Piqueras, J. A., Espinosa-Fernández, L., García-Lopez, L. J., & Beidel, D. C. (2012). Validación del Inventario de ansiedad y fobia social-forma breve"(SPAI-B) en jóvenes adultos españoles [Validation of the Social Phobia and Anxiety Inventory - Brief Form (SPAI-B in young Spanish adults.]. *Behavioral Psychology/Psicología Conductual, 20*(3), 505-529.

Rivollier, F., Peyre, H., Hoertel, N., Blanco, C., Limosin, F., & Delorme, R. (2015). Sex differences in DSM-IV posttraumatic stress disorder symptoms expression using item response theory: a population-based study. *Journal of Disorders, 187*, 211–217. doi: 10.1016/j.jad.2015.07.047

Rubin, M., Evans, O., & Wilkinson, R. B. (2016). A Longitudinal Study of the Relations Among University Students' Subjective Social Status, Social Contact with University Friends, and Mental Health and Well-Being. *Journal of Social and Clinical Psychology*, *35*(9), 722-737. doi: 10.1521/jscp.2016.35.9.722

Tennant, A. & Pallant, J. (2007). DIF matters: a practical approach to test if Differential Item Functioning makes a difference. *Rasch Measurment Transaction 20*(4),1082–1084.

Turner, S. M., Beidel, D. C., Dancu, C. V., & Stanley, M. A. (1989). An Empirically Derived Inventory to Measure Social Fears and Anxiety. *Psychological Assessment, 1*(1), 35-40. doi: 10.1037//1040-3590.1.1.35

Wang, W. C. & Chen, C. T. (2005). Item parameter recovery, standard error estimates, and fit statistics of the WINSTEPS program for the family of Rasch models. *Educational and Psychological Measurement*, *65*(3), 376-404. doi: 10.1177/0013164404268673

Wright, B. & Linacre, M. (1994). Reasonable mean-square fit values. *Rasch Measurement Transactions, 8*(3), 370.

Wright, B. D. & Masters, G. N. (1982). *Rating scale analysis*. Chicago: Mesa Press.

Wright, B. D. & Stone, M. H. (2003). Five steps to science: Observing, scoring, measuring, analyzing, and applying. *Rasch Measurement Transactions*, *17*(1), 912-913.

Velásquez, C., Montgomery, U. W., Montero, L. V., Pomalaya, V. R., Dioses, C. A., Velásquez, N. C. A., … Reynoso, E. D. (2008). Bienestar psicológico, asertividad y rendimiento académico en estudiantes universitarios sanmarquinos [Psychological well-being, assertiveness and academic performance in San Marcos university students]. *Revista Investigación Psicológica, 11*(2), 139-152. doi: 10.15381/rinvp.v11i2.3845