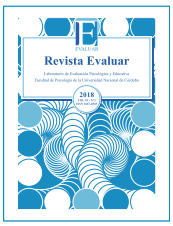
EVALUAR

2018, Vol. 18, No. 3
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



Advantages of the Rasch Model for Analysis and Interpretation of Attitudes: the Case of the Benevolent Sexism Subscale

Ventajas del Modelo de Rasch para el análisis e interpretación de actitudes: El caso de la subescala de sexismo benevolente

José Andrey Zamora-Araya * [1, 3], Vanessa Smith-Castro [2], Eiliana Montero-Rojas [2, 3], Tania Elena Moreira-Mora [4]

1 - Escuela de Matemáticas, Universidad Nacional, Heredia, Costa Rica.
2 - Instituto de Investigaciones Psicológicas, Universidad de Costa Rica, San José, Costa Rica.
3 - Escuela de Estadística, Universidad de Costa Rica, San José, Costa Rica.
4 - Instituto Tecnológico de Costa Rica, Cartago, Costa Rica, Departamento de Orientación y Psicología.

Recibido: 24/03/2018 **Revisado:** 10/05/2018 **Aceptado:** 28/05/2018

Abstract

This paper describes how the use of Rasch Analysis (RA), compared with the Classical Test Theory (CTT) and other Item Response Theory (IRT) approaches, could enhance the study and interpretation of attitudinal scales. This is illustrated with data from 197 students from the University of Costa Rica who answered the Benevolent Sexism (BS) Scale (Glick & Fiske, 1996). Besides providing estimations of the measure's specific accuracy at different levels of the construct, the RA, thanks to the person versus item map, allowed us to generate respondents' profiles describing particular aspects of the construct and according to their estimated scores in the scale. The analysis indicated that construct categories for participants with scores between [-0.30, 0.5] in the logit scale are the most accurately represented, with more items covering this interval, and reflecting the three aspects of the scale described by the theory. On the other hand, results showed less measurement accuracy for a considerable number of respondents with lower scores, suggesting the need for the development of additional items for that level of the scale. These evidences are discussed in light of the benefits of using the RA for the understanding and interpretations of respondents' scores in attitudinal scales, according to the underlying theory.

Keywords: *Benevolent Sexism Scale, Classical Test Theory, Rasch Analysis, Extended Rasch Model, attitude scales, psychometric analysis*

Resumen

Este artículo describe cómo el uso del Análisis de Rasch (AR), comparado con la Teoría Clásica de los Tests (TCT) y otros modelos de Teoría de Respuesta a los Ítems (TRI), puede mejorar el estudio y la interpretación de escalas actitudinales. Esto se ilustra con datos de 197 estudiantes de la Universidad de Costa Rica que tomaron la escala de Sexismo Benevolente (SB; Glick & Fiske, 1996). Además de proveer estimaciones específicas de la precisión de la medición en diferentes niveles del constructo, el Análisis de Rasch, gracias al mapa de personas versus ítems, permitió generar perfiles de los participantes en términos de aspectos particulares del constructo y de acuerdo con sus puntajes estimados en la escala. El análisis indicó que las categorías del constructo para participantes con puntajes entre [-0.30, 0.5] en la escala logit son las que están mejor representadas, con más ítems que cubren este intervalo y reflejan los tres aspectos de la escala descritos en la teoría. Por otro lado, los resultados mostraron menos precisión en la medición para un considerable número de participantes con puntajes más bajos, lo que sugiere la necesidad de desarrollar ítems adicionales para este nivel de la escala. Estas evidencias son discutidas a la luz de los beneficios de usar el AR para el entendimiento e interpretación de puntajes en escalas de actitud, de acuerdo con la teoría subyacente.

Palabras Clave: *Escala de Sexismo Benevolente, Teoría Clásica de los Tests, Análisis de Rasch, Modelo Extendido de Rasch, escalas de actitud, análisis psicométrico*

How to cite: Zamora-Araya, J. A., Smith-Castro, V., Montero-Rojas, E., & Moreira-Mora, T. E. (2018). Advantages of the Rasch Model for analysis and interpretation of attitudes: The case of the Benevolent Sexism Subscale. *Revista Evaluar, 18*(3), 1-13. Available at https://revistas.unc.edu.ar/index.php/revaluar
*** Correspondence to:** José Andrey Zamora Araya, PO Box. 86-3000, Heredia, Costa Rica. Tel: (506) 2562-6029. E-mail: jzamo@una.ac.cr
Authors' note This research was supported by a grant of the Council of Rectors of Costa Rican Universities to the Project "Nueva formas de medir viejas ideologías: el caso del sexismo y sus implicaciones en el ámbito académico (New Forms of Measuring Old Ideologies: the Case of Sexism and its Implications in the Academic Domain)," of the University of Costa Rica, the National University and the Costa Rica Institute of Technology.

Introduction

For many years, researchers in the field of Social Psychology have commonly applied the Classical Test Theory (CTT) approach in order to gather evidence on the reliability of its measures. This approach is based on the assumption that the value of an attribute is represented by an observed score, which is the sum of a true score (error-free) and the measuring error. Although CTT can provide important evidence of the accuracy of measuring instruments, several new psychometric tools might complement or even replace this approach in order to collect more accurate evidence to support the inferences made about the meaning and interpretation of scores (Muñiz, 2017). According to Bond and Fox (2001), one of such tools is the Rasch Analysis (RA), through which trait levels (the probability of a correct response or the probability of endorsing any option on each item) are modeled as a mathematical function of the difference between the person and the item parameters (Prieto-Adanes & Dias-Velasco, 2003).

This study describes the results of applying both, CTT and RA, to test the measurement properties of the Benevolent Sexism Scale (BS), one of the two subscales of the Ambivalent Sexism Inventory (ASI) developed by Glick and Fiske (1996). Our goal is to illustrate, with this subscale, the benefits of using RA to attain a better understanding of the strengths and weaknesses of instruments in the affective domain.

The Rasch Model: Characteristics and Advantages over CTT and other IRT Models

As pointed out before, most psychometric tests have been analyzed using the CTT. This model assumes that *X*, the observed score in the test, is a linear combination of two quantities, the true score (*T*) and the measurement error (*E*): X = T + E (Muñiz, 2017).

One of the limitations of the CTT is the assumption that *E* is constant across true score values, i.e. the error associated to each examinee is the same, no matter what his/her *X* is. Even intuitively, this assumption seems empirically unlikely. If the test items are endorsed by the majority of the respondents, it is fair to conclude that scores in the higher level of the trait will be estimated with less precision (more error) than scores in the lower end of the trait. On the other hand, if the items are endorsed by only a few respondents, scores in the lower level of the trait will be estimated with less precision (more error) than the scores in the higher end of the trait. Therefore, when applying CTT, it is not possible to provide different precision estimates for the different levels of the construct being measured. However, researchers and practitioners often need to measure certain levels of the construct with more precision, depending on their particular purposes and applications.

Another fundamental shortcoming of CTT lays in the fact that the model does not allow for descriptive interpretations of the meaning of each particular score. This limitation was first noticed in the educational measurement community (Wilson, 2004), which traditionally criticized the CTT approach for not addressing the need of knowing what students can and cannot achieve according to their scores in the tests. Social psychologists could also benefit from the possibility of attaching specific meanings to each particular score in their scales in terms of the construct being measured.

To overcome the limitations of CTT, a family of models known as IRT (Item Response Theory) were proposed around the second half of the twentieth century (Hambleton & Swaminathan, 2013). These mathematical models attempt to describe the respondents' behavior based on their an-

swer to each item. In general, the logistic function is used to estimate the model, with three different formulations: 1PL is the One Parameter Logistic model, 2PL is the Two Parameter Logistic model, and, 3PL is the Three Parameter Logistic model. The difference between these models lies on the number of parameters needed for their definition. In the 1PL model only the item difficulty, *b*, is estimated, along with the examinee's ability; in the 2PL model the item discrimination, *a*, is also estimated; and in the 3PL model, a guessing parameter, *c*, is estimated as well. 1PL is obtained when the item discrimination is assumed constant for all the items and the guessing parameter is assumed to be zero; on the other hand, the 2PL is obtained when only the guessing parameter is set to zero. Thus, both models are special cases of the IRT 3PL model.

The 1PL model is also known as the Rasch Model in honor of the Danish mathematician Georg Rasch, who in the 1960's described the special properties that only this model possesses (Olsen, 2003), making it particularly useful, and very attractive for applied test users who are interested in knowing what their instruments allow them to infer in terms of substantive interpretations (Rasch, 1980). Its mathematical formula relates the probability of the outcome (response) to the level of the respondant in the construct under measurement, and, the item difficulty. *Difficulty* is a term also used for tests in the affective domain that describes how low or high is the mean score for a specific item. In this case it can be also described as *endorsability*.

The Rasch model, in its original form, for dichotomous items is written as follows:

(1)

$$P\left(X_{ij}=1\right)=\frac{e^{(\theta-b_i)}}{1+e^{(\theta-b_i)}}$$

Where,

P (X_{ij} = 1): Probability that a specific person *j* answers correctly to the item *i*, and 0 for any other case.

θ: Parameter that describes a specific level of the trait for a person *j*.

b_i: Parameter that describes the difficulty (endorsability) of the item *i*.

θ and *b* can take any value in the real domain and they are both in the logit scale.

This initial formulation describes the Rasch Model, referring to the dichotomous items in the cognitive domain (1 correct, 0 incorrect). However, later developments have shown that it can be easily extended to data from rating scales for instruments in the affective domain, such as traits estimated through the Likert scale (Carvalho, Primi, & Meyer, 2012). For example, suppose that we have an item with *m*+1 response options. In this case, each of the m first options are described by the following expression:

(2)

$$P_{ik}(\theta)=\frac{e^{\sum_{k=0}^{h}(\theta-b_{ik})}}{\sum_{h=0}^{m}e^{\sum_{k=0}^{h}(\theta-b_{ik})}}$$

Where

$h = 0,1,\ldots, m$

$P_{ik}(\theta)$: indicates the probability of a subject with a specific θ score to endorse category *k* in item *i*.

b_{ik} is the endorsability parameter for item *i* in the *k* category.

m+1 is the total number of response categories.

Note that the probability for endorsement of the last category (i.e. reference category) is obtained when the examinee does not endorse any of the other *m* categories. In fact, these endorsement parameters estimated for affective scales are equivalent to the difficulty parameters estimated with dichotomous scales.

There is ample evidence that attribute the relative robustness of the Rasch Model to the deviation from the assumptions of equality of discrimination and zero guessing. In terms of robustness regarding these two specific assumptions, Muñiz, Rogers and Swaminathan (1989) found, by means of simulations, that estimations and fit indexes in the Rasch Model do not present great differences when there is guessing and variability in discrimination indexes.

Within the Rasch Model, as in the other IRT models, each particular estimated score has a specific estimation of its measurement error. Hence, it is possible to estimate how well the test's scores in the low, medium and high end of the scale might be. It also allows for the selection of the items that provide more precision (less errors) in pre-specified intervals of the trait under measurement. In other words, the measurement error is not the same for all examinees but it is a function of θ (Muñiz, 1997).

The specific advantage of the Rasch Model over other IRT models is that the estimated values for person and items are in the same scale of latent units (logits). This property is called *conjoint measurement*, which can be used to generate criterion-referenced interpretations in terms of qualitative descriptions of what the examinee can or cannot do (or what the examinee agrees or does not agree to do). This is possible thanks to the person-by-item map. Thus, the interpretation of scores in the Rasch Model is not based on group norms (as typically done in CTT), but it can be done in terms of item content and processes in which the examinee has a low or high probability of answering correctly (or has a low or high probability of endorsing). This trait provides the Rasch Model with a great diagnostic power.

Goodness of Fit Criteria

As Bond and Fox (2001) point out, before interpreting results in the Rasch Model it is necessary to check if the data adjusts reasonably to the model. There are several statistical measures of fit that can be used in this context, but one of the most widely used is called INFIT, which is an internal fit indicator corresponding to the residuals' weighted quadratic mean. Since items and persons are measured along the same scale, INFIT can be calculated for both of them.

The formula to obtain this measure is the following:

(3)

$$INFIT = \sum \frac{z^2_{vi} W_{vi}}{N}$$

Where each observation, (item endorsability or person's level in the construct) is weighted by its individual variance.

INFIT gives more importance to the examinees or items whose trait level is located near the item difficulty or person ability. Thus, at the examinee level, the INFIT indicator will attach more weight to items with difficulties (agreeability or endorsability) near the examinee's score. Conversely, at the item level, INFIT will give more weight to persons' ability estimates that are near the item difficulty (Bond & Fox, 2001).

Smith, Schumaker and Bush (as cited by Prieto-Adanes & Dias-Velasco, 2003) recommend different intervals to evaluate INFIT depending on sample size. Thus, INFIT values higher than 1.3 indicate lack of fit in samples with less than 500 subjects, 1.2 is the threshold value for samples between 500 and 1000 subjects, and 1.1 is the threshold value for samples with more than 1000 subjects.

Rasch models have been employed to test psychometric properties of tests intended to measure performance, abilities and competences.

Their employment on instruments for measuring attitudes, motivations, interests, values, subjective appreciations or psychological traits (the so called affective domain) is less frequent. The present study illustrates, using the Benevolent Sexism Subscale, how the Rasch Model improves the analyses and interpretations of these types of scales, compared to the CTT.

Benevolent Sexism: Conceptualization

Over the past 30 years, research on sexism against women has provided compelling evidence of the pervasiveness of anti-female biases in our societies. *Sexism* has been traditionally defined as the endorsement of discriminatory or prejudicial beliefs and feelings based on sex, usually linked with stereotypical conceptions of the sexes and the adoption of a traditional gender-role ideology (Moya & Expósito, 2001).

Currently, considerable attention has been paid to contemporary forms of sexism against women in the light of two observations: First, in the current cultural climate it is unlikely that respondents will openly endorse prejudicial attitudes toward women (Campbell, Schellenberg, & Senn, 1997). Second, given the particular intimate relationship between men and women, sexism against women does not always reflect open hostility, but rather a profound ambivalence (Glick & Fiske, 1996).

In trying to capture the complexity of contemporary forms of sexism, several researchers have conceptualized it in different ways. For instance, Glick and Fiske (1996) describe sexism as a multidimensional construct that involves both, hostile and benevolent attitudes toward women. *Hostile sexism* is characterized as antipathy and derogatory attitudes, as in the classical definition of prejudice, while *benevolent sexism* is defined as a set of subjectively positive attitudes that are sexist in terms of typecasting women in restricted roles.

Consequently, Glick and Fiske (1996) developed a scale attempting to measure this construct accurately and reliably: the Ambivalent Sexism Inventory (ASI), which comprises two subscales: the Hostile Sexism Scale (HS) and the Benevolent Sexism Scale (BS). A detailed description of the instrument is presented in the method section.

In the present study we focus specifically on BS because of the unique characteristics of the construct and its relevance for understanding contemporary forms of sexism. Despite of the subjectively positive content of the scale, it reflects sexism by justifying traditional gender roles and masculine dominance (e.g., the man as the provider and woman as his dependent). Interestingly, benevolent sexist attitudes are not always recognized as such by respondents, who tend to endorse BS items more strongly than HS items. Moreover, participants who endorse benevolently sexist beliefs are more likely to endorse other gender-traditional attitudes, including hostile sexism, unaware of the fact that they are endorsing two complementary aspects of the same sexist ideology.

Additionally, BS is harmful for women by itself, not only because of its relationship with HS. Data show that men who endorse BS are more likely to blame a female victim of rape if she has "infringed" traditional gender role expectations (Viki, Abrams, & Masser, 2004); and women who endorse BS are more likely to accept an ostensibly protective and restrictive male as a romantic partner, even if it implies a constraint to their career aspirations (Moya, Glick, Expósito, De Lemus, & Hart, 2007). In sum, as Glick and Fiske (1996) point out, the BS Scale measures an aspect of sexism with important consequences for women that many other instruments might overlook.

To the best of our knowledge, no study describing the adaptation or validation of this specific subscale using Rasch Analyses has been published so far. Therefore, analyzing it with this approach could be useful for illustrating the benefits of the Rasch Model when it comes to a deeper understanding of psychometric properties of scales in the affective domain.

Method

Participants

Analyses were run on a random cluster sample of 197 students from the University of Costa Rica, the National University of Costa Rica, and the Costa Rica Institute of Technology. These are the main State universities of the Country, located in the Metropolitan Area of the Central Valley of Costa Rica. One hundred and sixty six (84.3% of the sample) were women. The mean age was 21.69 years (SD = 3.67 years). Inclusion criteria were: a) being an active student of introductory Humanities and Math courses at these universities, and b) voluntarily participating in the study.

Instruments

The paper-pencil questionnaire contained a brief demographic section, along with several measures of attitudes toward women, including a Latin American adaptation of the ASI (Cárdenas, Lay, González, Calderón, & Alegría, 2010). The 22-Item ASI is made up of two subscales: HS, which basically matches the old sexism conceptualization, and BS, reflecting women as delicate creatures, confined to limited roles. Examples of HS items are *Women seek to gain power by getting control over men* and *Women exaggerate problems they have at work*. Examples of BS items are *Many women have a quality of purity that few men possess* and *Women should be cherished and protected by men*. Items are rated in a 5-point Likert scale.

Glick and Fiske (1996) reported Cronbach's alpha coefficients for the overall scale ranging from .80 to .90. For the HS subscale, alphas have been ranged from .80 to .90, whereas for BS subscale alphas are lower, ranging from .70 to .85. Their validity studies yielded significant correlations between ASI, specially the HS, and other measures of sexism, racism and gender biases. Further analytic evidence supports the idea that the ASI scores show two correlated yet distinct primary dimensions: hostile and benevolent sexism (Glick & Fiske, 1996).

Other authors have also provided evidence of the reliability and validity of ASI-scores not only among adults (Becker & Wagner, 2009; Cárdenas et. al, 2010), but also among adolescents (De Lemus, Moya, & Glick, 2010; Etchezahar & Ungaretti, 2014) in different social, cultural and linguistic contexts (Chen, Fiske, & Lee, 2009; Rodríguez-Castro, Lameiras-Fernández, & Carrera-Fernández, 2009; Rodríguez & Magalhães, 2013; Sakalli-Uğurlu & Glick, 2003). A general description of the psychometric properties of the measure from the CTT approach can be found in Fiske & North (2014). The complete scale and scoring instructions are available in Glick and Fiske (1996).

Procedures

Questionnaires were group administered to the students in their classrooms. Following the guidelines of the Institutional Revision Board (IRB) of the University of Costa Rica, respondents were informed about the purpose of the study, that their participation was voluntary, that

no reward would be given and that the personal information will remain confidential.

Analyses

To test the psychometric properties of BS from the perspective of the CTT, means, standard deviations, and item-total correlations were calculated for all items, as well as Cronbach's alpha coefficient and standard error of measurement for the total scale using SPSS 21 (IBM Corporation, 2012). RA comprised persons and items fit analyses, using INFIT statistics. INFIT values between 0.5 and 2.0 were considered acceptable for respondents' fit (Linacre, 2002), whereas values ranging from 0.7 to 1.3 were considered satisfactory for items' fit (Prieto & Delgado, 2003). Secondly, the Extended Rasch Model was estimated using joint maximum likelihood using the WINSTEPS 3.72.3, including respondents' scores and items' endorsabilities (difficulties), reliabilities, measurement errors, as well as the person-item map.

Data Preparation and Preliminary Analyses

In preparation for the main analyses, data was screened to detect major problems with asymmetrical distributions, missing values and outliers. Since diagnostic analyses revealed no major issues in this regard, all items were retained for further analyses. Only respondents who satisfactorily fitted the Extended Rasch Model (N = 197), were employed for comparison and contrast purposes.

Results

Table 1 shows some of the principal psychometric properties of BS obtained by means of employing the CTT and RA. From the CTT perspective, data shows a Cronbach's alpha of .74, item means ranging from 1.49 to 2.67 (in a scale from 1 to 5), and item-total correlations from .30 to .51, with exception of item BS6, which shows an unacceptable item-total correlation of .04. Regarding RA, data revealed an average respondents' reliability of .68, which means that if the same group of participants were to answer to another set of items drawn from the same hypothetical item universe, the estimated correlation between the two estimations of the construct would be approximately .68. RA also showed item endorsability parameters ranging from -0.81 to 0.75, in the logit scale. Only one item, BS6, showed an unacceptable INFIT value of 2.13. Therefore, item BS6 was left out for the subsequent analyses. Standard errors of measurement ranged from .06 to .09, as shown in Figure 1, and in the persons vs. items map in Figure 2.

Finally, RA can be used to generate criterion-referenced interpretations about respondents' attitudes. For BS scale, the person vs. item-map allows creating categories of responses corresponding to different levels of endorsability. This is illustrated by the analysis of item content shown in Table 2.

Discussion

The purpose of the present study is to illustrate the benefits of using RA for the analysis and interpretation of attitudinal scales by applying RA and CTT procedures to the Benevolent Sexism Scale.

There are some similarities in the information provided by both approaches, since both models yielded statistics indicating poor psychometric quality for item BS6, in one case because it shows an item-total correlation of .04, in the

Table 1
Statistical properties for BS under CTT and RA.

	CTT				RA		
	Cronbach's alpha = .745				Person Reliability = .68		
Item	**Difficulty**	**Standard deviation**	**Standard error of measurement**	**Item-total correlation**	**Logit**	**Standard error of measurement**	**INFIT**
BS1	2.05	1.324		.470	0.14	0.070	0.96
BS3	2.67	1.300		.304	-0.33	0.060	0.98
BS6	3.05	1.743		.038	-0.59	0.060	2.13
BS8	2.15	1.285		.508	0.05	0.070	0.80
BS9	3.37	1.451		.466	-0.81	0.060	0.92
BS12	2.21	1.352		.556	0.01	0.060	0.80
BS13	1.49	0.932		.509	0.75	0.090	0.84
BS17	1.92	1.254		.457	0.25	0.070	0.96
BS19	2.06	1.183		.459	0.13	0.070	0.79
BS20	1.60	1.025		.484	0.60	0.080	0.87
BS22	2.49	1.375		.407	-0.20	0.060	0.94
TOTAL	25.06	7.638	3.857				

Note. BS = Benevolent Sexism. Item numbering corresponds to Glick and Fiske (1996).

other because it presents an INFIT of 2.13. As highlighted before, both approaches emphasized the need of removing this item for any subsequent analyses.

In other aspects, however, both models offered different results. CTT assumes a constant measurement error, which in this case was equal to 3.857 (in a total scale ranging from 11 to 55), i.e., regardless of the construct level, all items are assumed to provide the same precision. On the other hand, RA relaxes this limiting assumption, estimating specific measurement errors for both respondents and items. Thus, in this case, measurement errors varied along different levels of the construct, being more accurate those near 0 in the logit scale, which is centered on the item average endorsability (see Figure 1).

CTT and RA also rendered different results regarding the total scores (i.e. the estimated construct level for each respondent). The Cronbach's alpha value of .745 suggests an acceptable internal consistency measure for research purposes. On the contrary, the estimated person reliability of 0.68 is clearly not satisfactory even for research purposes.

As it was previously mentioned, a very valuable feature of RA is its conjoint measurement property. It means that estimations for respondents' scores and items' means (i.e., endorsabilities) are calculated in the same logit scale. In this regard, Figure 2 depicts this useful property that CTT does not offer, showing that there are very few items measuring construct levels lower than 0. In addition, measurement accuracy increases as

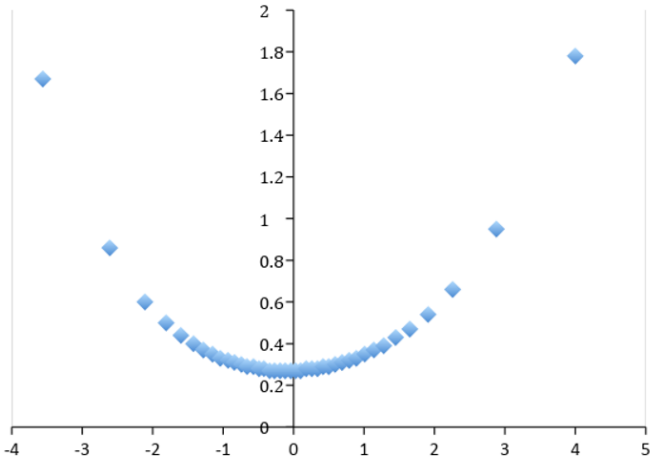


Figure 1
Measurement errors for BS in Rasch Analysis.

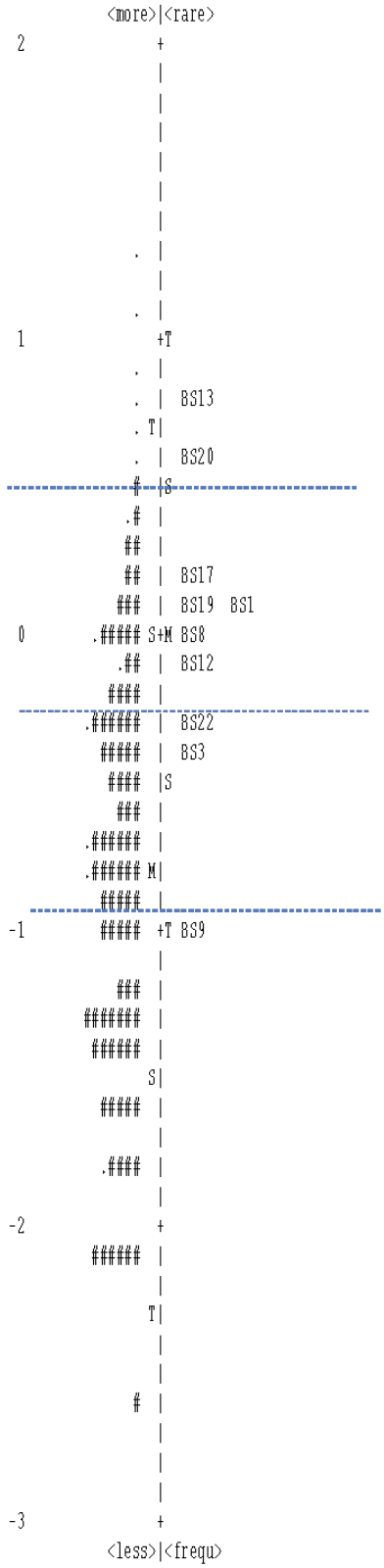
Note. Each dot represents the specific measurement error associated to each score in the construct. X axis: the level of the construct in logit units. Y axis: measurement error in logit units. 0 = Mean difficulty (endorsability) of the items.

the estimated scores reach values close to 0. This means that, for a considerable number of people, the construct cannot be accurately measured with the BS scale; in particular for those who are more likely to disagree with item content.

Taking into account the theoretical background of the BS, expert judgments and respondents' scores distribution, we noticed that only one item (BS9) represented the lower range of the BS scale [-2.61, -0.83] for this group. The content of this item reflects a mild kind of protective paternalism towards women, which might be seen as an inoffensive form of modern-day chivalry. Because these kind of benevolent sexist attitudes seem positive, participants might not recognize these beliefs as a form of gender-based prejudice, therefore this kind of items are more likely to be endorsed.

Participants in the next level [-0.83, -0.30], do not only endorsed paternalistic attitudes in

Figure 2
Persons vs. items map for the BS Scale.



Note. Each"#" is 2, each "." is 1. BS = Benevolent Sexism.

Table 2
Subjects' BS profiles.

Score interval	Description	Content
[0.5, 1.29]	BS13. Women are incomplete without men (0.81)	Heterosexual intimacy
	BS20. Men should be willing to sacrifice their own well-being in order to provide financially for the women in their lives (0.64)	Protective paternalism.
[-0.30, 0.5]	BS17 A good woman should be set on a pedestal by her man (0.24)	Protective paternalism
	BS1 No matter how accomplished he is, a man is not truly complete as a person unless he has the love of a woman (0.09)	Heterosexual intimacy
	BS19 Women, compared to men, tend to have a superior moral sensibility (0.09)	Complementary gender differentiation
	BS8 Many women have a quality of purity that few men possess (0.00)	Complementary gender differentiation
	BS12 Every man ought to have a woman whom he adores (-0.06)	Heterosexual intimacy
[-0.83, -0.30]	BS22 Women, as compared to men, tend to have a more refined sense of culture and good taste (-0.32)	Complementary gender differentiation
	BS3 In a disaster, women ought to be rescued before men (-0.44)	Protective paternalism.
[-2.61, -0.83]	BS9 Women should be cherished and protected by men (-1.05)	Protective paternalism.

Note. BS = Benevolent Sexism. In the second column, values within parentheses are item difficulties.

form of chivalry, but also stereotypical complementary gender differentiation; that is, participants with this level of sexism tended to endorse the idea that women are delicate creatures, and that they therefore need to be protected.

Construct levels of those participants with scores between [-0.30, 0.5] are the most accurately represented with this instrument. There are five items covering this interval, reflecting the three aspects of BS described by the theory, i.e. protective paternalism, complementary gender differentiation and heterosexual intimacy. At this level, participants not only endorsed mild forms of modern-day chivalry and the notion that women are delicate creatures in need of protection, but also the idea that men are incomplete without women, reflecting heterosexual intimacy, i.e. the belief that romantic intimacy is necessary to complete a man, but also that women are incomplete without men.

Finally, the most *difficult* items; i.e., those which are *more difficult* to be endorsed by participants, turned out to be a combination of an extreme form of protective paternalism (BS20), and a plain statement that a woman is incomplete without a man by her side (BS13). Participants with this level of benevolent sexism are more willing to accept that women are so defenseless that men should sacrifice themselves in order to protect them, reflecting not only the superiority of men over women, but also undermining the notion of women as competent and independent agents.

Notice that in both the lower and the higher levels of the construct, the measurement was less accurate; since not all three components of BS were present along the continuum and because of the reduced number of items, which resulted in a less precise measurement (see Figure 1).

Conclusion

Our data showed that the Extended Rasch Model is a useful tool for testing psychometric aspects of scales in the attitudinal domain such as the Benevolent Sexism Subscale. It also allows researchers and practitioners to generate meaningful interpretations about the construct being measured. CTT, although useful for some purposes, is more restrictive; presenting important shortcomings that RA helps to overcome. This paper illustrates several valuable features of RA, as fit statistics for both persons and items, and specific estimations for measurement error at different levels of the construct. More importantly, the conjoint measurement property provides the Extended Rasch Model with a particular advantage over other IRT models, allowing researchers to generate respondents' profiles and criterion-referenced interpretations.

This is particularly important for measures in the attitudinal domain. While CTT and other IRT models allow researchers to compute a score on BS, reflecting a global view of participants' BS (low or high) levels; RA allows researchers to understand which items are more probable to be endorsed by participants with different levels of BS, enhancing knowledge about the meaning of low and high scores in the measure. Using this tool, we can better understand how benevolent sexist attitudes toward women are constituted and organized, providing a deeper comprehension of contemporary sexism in our societies, which will also contribute to the development of educational programs and community interventions to foster social equity and justice.

References

Becker, J. C., & Wagner, U. (2009). Doing gender differently: The interplay of strength of gender identification and content of gender identity in predicting women's endorsement of sexist beliefs. *European Journal of Social Psychology, 39*(4), 487-508. doi: 10.1002/ejsp.551

Bond, T. G., & Fox, C. M. (2001). *Applying the Rasch model: Fundamental measurement in the human sciences.* Mahwah, NJ: Lawrence Erlbaum Associates.

Campbell, B., Schellenberg, E. G., & Senn, C. Y. (1997). Evaluating measures of contemporary sexism. *Psychology of Women Quarterly, 21*(1), 89-101. doi: 10.1111/j.1471-6402.1997.tb00102.x

Cárdenas, M., Lay, S. L., González, C., Calderón, C., & Alegría, I. (2010). Inventario de sexismo ambivalente: Adaptación, validación y relación con variables psicosociales. *Revista Salud y Sociedad, 1*(2), 125-135. doi: 10.22199/s07187475.2010.0002.00006

Carvalho, L., Primi, R., & Meyer, G. J. (2012). Aplicação do modelo de Rasch na medida de transtornos da personalidade. *Trends in Psychiatry and Psy-*

chotherapy, 34(2), 101-109. doi: 10.1590/S2237-60892012000200009

Chen, Z., Fiske, S. T., & Lee, T. L. (2009). Ambivalent sexism and power-related gender-role ideology in marriage. *Sex Roles, 60*(11-12), 765-778. doi: 10.1007/s11199-009-9585-9

De Lemus, S., Moya, M., & Glick, P. (2010). When contact correlates with prejudice: Adolescents' romantic relationship experience predicts greater benevolent sexism in boys and hostile sexism in girls. *Sex Roles, 63*, 214-225. doi: 10.1007/s11199-010-9786-2

Etchezahar, E., & Ungaretti, J. (2014). Woman stereotypes and ambivalent sexism in a sample of adolescents from Buenos Aires. *Journal of Behavior, Health & Social Issues, 6*(2), 87-94.

Fiske, S. T., & North, M. S. (2014). Measures of stereotyping and prejudice: Barometers of bias. In G. Boyle & D. Saklofske (Eds.), *Measures of Personality & Social Psychological Constructs* (pp. 684-718). Boston, MA: Elsevier Academic Press.

Glick, P., & Fiske, S. T. (1996). The Ambivalent Sexism Inventory: Differentiating hostile and benevolent sexism. *Journal of Personality and Social Psychology, 70*(3), 491-512. doi: 10.1037/0022-3514.70.3.491

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications.* Berlín: Springer Science & Business Media.

IBM Corporation (2012). IBM SPSS Statistics for Windows (Version 21.0) [computer software]. Armonk, NY: IBM

Linacre, J. M. (2002). What do infit and outfit, mean-square and standardized mean? *Rasch Measurement Transactions, 16*(2), 878. Retrieved from https://www.rasch.org/rmt/index.htm

Moya, M., & Expósito, F. (2001). Nuevas formas, viejos intereses: Neosexismo en varones españoles. *Psicothema, 13*(4), 643-649. Retrieved from http://www.psicothema.com

Moya, M., Glick, P., Expósito, F., De Lemus, S., & Hart, J. (2007). It's for your own good: Benevolent sexism and women's reactions to protectively justified restrictions. *Personality and Social Psychology Bulletin, 33*(10), 1421-1434. doi: 10.1177/0146167207304790

Muñiz, J. (1997). *Introducción a la teoría de respuesta a los ítems.* Madrid, España: Pirámide.

Muñiz, J. (2017). *Teoría Clásica de los Tests.* Madrid, España: Pirámide.

Muñiz, J., Rogers, J., & Swaminathan, H. (1989). Robustez de las estimaciones de modelo de Rash en presencia de aciertos al azar y discriminación variable de los ítems. *Anuario de Psicología, 43*(4), 82-97. Retrieved from https://www.raco.cat/index.php/AnuarioPsicologia

Olsen, L. W. (2003). Essays on Georg Rasch and his contributions to statistics (Unpublished doctoral dissertation, University of Copenhagen, Institute of Economics) Retrieved from http://www.rasch.org/olsen.pdf

Prieto, G., & Delgado, A. R. (2003). Análisis de un test mediante el modelo de Rasch. *Psicothema, 15*(1), 94-100. Retrieved from http://www.psicothema.com

Prieto-Adanes, G., & Dias-Velasco, A. (2003). Uso del modelo de Rasch para poner en la misma escala las puntuaciones de distintos tests. *Actualidades en Psicología, 19*(106), 5-23. doi: 10.15517/ap.v19i106.43

Rasch, G. (1980). *Probabilistic models for some intelligence and attainment tests.* Chicago: The University of Chicago Press.

Rodríguez-Castro, Y., Lameiras-Fernández, M., & Carrera-Fernández, M. V. (2009). Validación de la versión reducida de las Escalas ASI y AMI en una muestra de estudiantes españoles. *Psicogente, 12*(22), 284-295. Retrieved from http://revistas.unisimon.edu.co/index.php/psicogente

Rodríguez, Y., & Magalhães, M. J. (2013). El sexismo moderno en estudiantes universitarios/as portugueses/as. *Revista Interdisciplinar de Ciencias Sociales y Humanas, 1*(2), 113-121. Retrieved from http://independent.academia.edu/revistaagir

Sakalli-Uğurlu, N., & Glick, P. (2003). Ambivalent sexism and attitudes toward women who engage in premarital sex in Turkey. *The Journal of Sex Research, 40*(3),

296-302. doi: 10.1080/00224490309552194

Viki, G. T., Abrams, D., & Masser, B. (2004). Evaluating stranger and acquaintance rape: The role of benevolent sexism in perpetrator blame and recommended sentence length. *Law and Human Behavior, 28*(3), 295-303. doi: 10.1023/b:lahu.0000029140.72880.69

Wilson, M. (2004). *Constructing measures: An item response modeling approach.* New York, NY: Routledge. doi: 10.4324/9781410611697