Applying a Rasch model to the Broad Autism Phenotype Questionnaire: item severity analysis and differential test functioning of the English and Spanish versions

Marta Godoy-Giménez[1] · Antonio González-Rodríguez[1] · Angeles F. Estévez[1,2] · Fernando Cañadas[1,2] · Pablo Sayans-Jiménez[1]

Abstract

The broad autism phenotype (BAP) represents the elevated but nonclinical levels of autism-related expressions expanding them beyond the disorder threshold towards the general population and it is assessed worldwide using the Broad Autism Phenotype Questionnaire. In this study, we applied a Rasch item response approach to the Spanish and the English BAPQ versions. This approach allows us to test the proper functioning of each version, estimate which BAP behaviours are more likeable at different levels of severity, and study whether BAPQ item's severities and orders could be considered equivalent between both versions. A Spanish community sample of 970 participants and an English of 533 ones completed either the Spanish or the English version of the BAPQ. The results revealed a lack of unidimensionality of the Pragmatic Language subscale in both tests and the Rigid subscale in the English version. Both Aloof subscale versions and the Spanish Rigid demonstrated adequate properties but with several items showing differential functioning in the case of Aloof. We conclude by highlighting the necessity of BAP measurements paralleling the current autism spectrum disorder structure, following the severity-dimensional conceptualization (behaviours are continuously distributed according to their severity), and suitable for being adapted to different languages.

The autism spectrum hypothesis proposes that autistic behaviours and personality expressions would be presented in clinical populations and within and along with the general population in different severity (Hoekstra et al., 2008). Inside this spectrum, the broad autism phenotype (BAP) is defined by elevated but nonclinical levels of autism-related symptoms expanding them beyond the threshold of the disorder towards the general population levels (De Groot & Van Strien, 2017). Thus, the BAP expressions inserted in the autism spectrum would contribute to understanding the disorder and its developmental trajectories (Landry & Chouinard, 2016). Also, the BAP opens the possibilities of carrying out studies with larger and more heterogeneous samples and, so, entailing a better representation of the population and guaranteeing greater statistical power (Mitchell & Jolley, 2013).

When studying the BAP, the Broad Autism Phenotype Questionnaire (BAPQ; Hurley et al., 2007) has been extendedly applied over different countries and cultures (e.g., China, Shi et al., 2015; Israel, Seidman et al., 2012). It was designed to measure the BAP in parents of children diagnosed with autism spectrum disorder (ASD; APA, 2013) and it has been subsequently applied to broader community samples (e.g.,Álvarez-Couto et al., 2021; Godoy-Giménez et al., 2018; Jakobson et al., 2018; Jamil et al., 2017; Morrison et al., 2018; Stojković et al., 2018). It represents the BAP original structure: Aloof personality, Rigid personality, and Pragmatic Language impairment (Piven et al., 1997a, b) which paralleled autism disorder domains of impairment of the prior DSM-IV definition (APA, 1994).

✉ Fernando Cañadas
jcanadas@ual.es

✉ Pablo Sayans-Jiménez
psj242@ual.es

1 Department of Psychology, University of Almería, 04120 Almería, Spain

2 Cernep Research Centre, University of Almería, 04120 Almería, Spain

While research in psychometric properties of the BAPQ has provided validity evidence on the BAPQ scores' inferences (based on their relations with other variables, Sasson et al., 2013b; on test-criterion relationships, Broderick et al., 2015; on differential scores among groups, Shi et al., 2015; and on comparing BAPQ versions, Sasson et al., 2014), little is known about BAPQ reliability at the different levels of the severity continuum, item severities or the adequacy of response categories. According to the autism spectrum hypothesis, understanding that autism expressions are spread somehow across the population implies that the items of the BAPQ could be scaled according to their severity. It would imply that the items conforming to each of the three BAPQ subscales should express different degrees of severity to grasp the essence of the spectrum of autism (at subclinical levels in the case of the BAP) and, at the same time, that every person may be located at some point inside this continuum even those who express mild severity BAP behaviours.

Item response theory (IRT) become of great relevance in this respect as they could bring many advantages in the study of BAPQ psychometric properties (e.g., Bond & Fox, 2015) providing, among others: (i) both interval-level scaling of persons and precision estimations for each severity level (essential for interpreting BAPQ's scores according to their severity), (ii) test-independent scores, useful for comparing BAP levels measured with different tests, or (iii) conjoint scaling of items and respondents along the continuum and for inferring which behaviours are more likeable at different levels of severity. This conjoint estimation will also allow inferring the amount of BAP severity of each item (the magnitude of BAP that would be required to adhere to each expression, scaled in a severity continuum from above-average functional behaviours to severe autistic-associated behaviours which require substantial help) and will ease the discussion about the differential performance of the items in different groups (e.g., ASD relatives vs. the general population). Furthermore, the resulting severity order can be used as an additional source of validity evidence.

On the other hand, many authors have highlighted the importance of assessing differential test functioning (DTF) and differential item functioning (DIF) to assure that all the test adaptions preserve the equivalence of measurement (item hierarchies and/or severities) among cultures (Hambleton, 1994). Unlike other questionnaires that have analysed test stability between versions (the Social Responsiveness Scale—SRS; Bölte et al., 2008), the BAPQ DTF has never been tested across test adaptations. This is especially important considering that some social expressions linked to ASD vary their severity depending on the country (Kim, 2012).

This study aims to further analyse the psychometric properties of the BAPQ through IRT models. The importance of this study relies on the ASD dimensional approach in the DSM-V (APA, 2013) and the autism spectrum hypothesis (De Groot & Van Strien, 2017) summed to the necessity of exploring whether the BAP operationalization targeted in the BAPQ would follow this severity-dimensional approach. That is, whether the items of the three subscales of the BAPQ could be graded in severity and displayed creating three hierarchies of severity in three different domains (corresponding to Aloof, Pragmatic Language, and Rigid subscales).

For those purposes, the original BAPQ (Hurley et al., 2007) and its Spanish version, the BAPQ-SP (Godoy-Giménez et al., 2018) will be applied to two community samples (one from the United Kingdom and one from Spain) and analysed using a Rasch rating scale model (RSM, Masters & Wright, 1984).

Secondly, we will test whether the items' hierarchies and severity of each BAPQ item remain stable in its Spanish version. DTF (equivalence in the functioning of sets of items) and DIF (the loss of item estimate invariance across subsamples of respondents; Bond & Fox, 2015) between both BAPQ versions will reveal that the members of different groups with the same level of BAP would answer in significantly different ways to the same items (e.g., Hambleton, 2006). Even if the severities of some autism-related characteristics could be affected by cultural idiosyncrasies, item invariance failures could also alert us to potential problems with the test (Bond & Fox, 2015). If so, we will consider whether these severity differences are due to cultural differences. This aim is also very important since the BAPQ is one of the most translated tests of the BAP (e.g., Bang et al., 2021 [Swedish version]; Godoy-Giménez et al., 2018 [Spanish version]; Shi et al., 2015 [Chinese version]) and no evidence has been documented about the invariance structure of the BAPQ either the functioning of its items across versions.

The following results would support the use of the original BAPQ and the BAPQ-SP: (i) an adequate item fit informing about the capability of the resulting item hierarchy to predict people patterns of responses (i.e., items statements should endorse progressively from the least to the most severe); (ii) separation indices indicating that it is possible to distinguish, at least, between two statistically different strata in the sample; (iii) item-person maps showing the BAPQ items are distributed covering the severity levels of the persons, and (iv) all the response categories being distinguishable for the respondents. We believe that our results will be in line with previous studies indicating adverse evidence based on the BAPQ internal structure (e.g., Lin et al., 2021; Stojković et al., 2018). Especially, a lack of unidimensionality of some BAPQ subscales is expected, mainly in the case of the Pragmatic Language or Rigid subscales (see Godoy-Giménez et al., 2018; Lin et al., 2021; Sasson et al., 2013a; Sharma and Bhushan, 2018; Stojković et al., 2018).

If results matched with expectations, it would be impossible to perform RSM in these subscales. Additionally, we expect to find some item severity equivalence among the BAPQ items when comparing the English (BAPQ-EN) and the Spanish (BAPQ-SP) versions. Finally, but not less important, in this study we also present the correction of two items of the Spanish BAPQ (items 4 and 23; Godoy-Giménez et al., 2018). Precedent Spanish version of item 4 does not completely represent the original content of the item in the BAPQ-EN (the Spanish item 4 is expressing the idea of following the track of the general conversation while the original item 4 in the BAPQ-EN is expressing keeping the track of your point in a conversation). For item 23, its Spanish version seems to express a preference for superficial conversations (an aloof prototypical behaviour; corresponding to a direct item) whereas in the BAPQ-EN this item reflects the idea of being good at casual or spontaneous conversations (reversed item). Alternative translations for these items are provided (see instruments section). After adjusting both translations, we will expect that they will not exhibit adverse item fit. A closer inspection of those items will be performed as we expect that alternative translations will work adequately and they will not generate any discrepancy in their severity estimation and comparisons between both BAPQ versions.

Method

Participants

Quota sampling was performed for having access to representative samples, regarding age and sex, of the Spanish and English populations. In the case of Spain, seven quotas by age and gender were calculated on a sample of 600 participants. Through incidental snowball, university students helped us to have access to a larger and general sample that involved themselves, relatives, friends, and acquaintances (undergraduate received two-course credits for collaborating). Later on, incidental sampling was conducted until reaching 970 Spanish participants (Social Networks and Spanish associations of autism took part). The English sample was recruited from an international company (https://www.gfk.com) and it consisted of 533 participants. We prepared eight quotas on a sample of 500 participants and we gave them to the company. The enterprise posted a notice with the study information in an online panel of respondents and participants who met the criteria were given access to the survey. Accessibility was restricted once a quota was completed. Participants received €6 for their collaboration. We took nationality as an exclusion criterion in both cases.

Sample sizes adequacy was calculated for item calibrations. Considering a test with 36 polytomous items, a sample of 500 would produce robust statistically stable measures (item calibrations or person measures stables within ± 1.0 logits, robust confidence) in adverse circumstances (Aziza et al.., 2020; Linacre, 2018; Wright & Stone, 1979). Furthermore, in the present study item separation indices were tested to confirm that both sample sizes were adequate to estimate the item hierarchies (as exposed hereafter, indices > 3 assume that the sample size is large enough to verify the item severity hierarchy; Linacre, 2018; and to find the same item placements if identical items are applied to another same-sized sample behaving similarly; Bond & Fox, 2015). Finally, for conducting comparisons between groups (testing differential item or test functioning) and for longer than two-item scale, a sample size of 200 participants per group is required for adequate power (> 80%; Scott et al., 2009).

Instruments

The BAPQ self-report form (Hurley et al., 2007) is a 36-item screening questionnaire that demands participants to answer how frequently a statement applies to them on a 6-point scale from 1 "Very rarely" to 6 "Very often". BAPQ items are grouped into three subscales: aloof personality, rigid personality, and pragmatic language problems (items conforming to each subscale and reversed items can be found in Hurley et al., 2007). The self-report BAPQ-SP (Godoy-Giménez et al., 2018) was applied in this study together with the original English self-report one. Alternatives translations of items 4 ("Me cuesta evitar irme por las ramas en una conversación") and 23[i] ("Se me da bien la charla insustancial (estar de cháchara)") were included this time.

Procedure

The data was compiled in a single session of approximately 30 min; both samples had equal conditions. A survey was created in the survey online administrator LimeSurvey (https://www.limesurvey.org/) and it includes the BAPQ together with another test out of scope of the present study and some sociodemographic questions always displayed at the end. The survey had two versions, the English-based language version and the Spanish one. A link drove each participant to an online and completely anonymous survey (links were different depending on the survey language). Personal links (anonymous tokens) were created for the English sample as the company required it to provide a refund for participants' collaboration. All the participants read the study instructions, the data treatment ethics, and expressed their consent to participate in the study and to the use of their data only for scientific purposes before starting. They were asked to read each question carefully and answer them sincerely. They were also warned that

Table 1 Item descriptive statistics and IRT parameters

Item	M (SD) EN	M (SD) SP	Sk EN	Sk SP	K EN	K SP	Measure EN	Measure SP	SE EN	SE SP	Infit EN	Infit SP	Outfit EN	Outfit SP	Item-total correlation EN	Item-total correlation SP	t-test t	t-test p	t-test d
Aloof																			
1[i]	3.30 (1.36)	2.74 (1.30)	0.03	0.39	-0.62	-0.60	-0.07	0.26	.04	.03	0.64	0.78	0.63	0.80	.81	.75	7.71	.00	.42
5	3.25 (1.29)	2.95 (1.39)	0.30	0.52	-0.53	-0.48	-0.02	0.02	.04	.03	1.20	1.04	1.26	1.11	.57	.70	-4.23	.00	.22
9[i]	3.40 (1.46)	3.09 (1.46)	0.06	0.26	-0.81	-0.81	-0.17	-0.13	.04	.03	0.69	0.80	0.68	0.81	.82	.79	-0.97	.00	.21
12[i]	2.91 (1.34)	3.04 (1.23)	0.31	0.26	-0.71	-0.91	0.34	-0.08	.05	.03	1.08	1.22	1.11	1.28	.65	.67	1.74	.08	.10
16[i]	3.14 (1.25)	3.67 (1.47)	0.03	-0.15	-0.60	-0.91	-0.47	-0.75	.04	.03	0.80	0.87	0.81	0.88	.76	.77	-0.46	.65	.39
18	3.25 (1.23)	2.83 (1.29)	0.24	0.65	-0.12	0.02	-0.01	0.16	.04	.03	1.45	1.25	1.65	1.41	.43	.59	-6.15	.00	.33
23[i]	3.54 (1.46)	2.83 (1.27)	0.01	0.66	-0.92	0.03	-0.30	0.16	.04	.03	1.17	1.32	1.18	1.36	.66	.57	-9.39	.00	.52
25[i]	3.54 (1.46)	3.11 (1.38)	0.01	0.29	-0.92	-0.65	-0.31	-0.16	.04	.03	0.84	0.86	0.83	0.86	.73	.75	-5.82	.00	.30
27	2.71 (1.32)	2.62 (1.42)	0.59	0.87	-0.07	-0.02	0.56	0.41	.05	.04	1.25	0.93	1.21	0.87	.60	.75	-1.27	.21	.07
28[i]	2.65 (1.23)	2.85 (1.39)	0.41	0.48	-0.55	-0.55	0.63	0.14	.05	.03	1.09	1.04	1.02	1.11	.62	.70	2.86	.00	.15
31	3.56 (1.41)	3.01 (1.32)	0.24	0.63	-0.68	-0.01	-0.33	-0.05	.04	.03	1.13	0.98	1.15	1.04	.65	.68	-7.35	.00	.40
36[i]	3.09 (1.42)	2.96 (1.25)	0.14	0.31	-0.81	-0.70	0.15	0.01	.04	.03	0.68	0.64	0.66	0.66	.82	.82	-1.74	.08	.10
PragLang																			
2	2.88 (1.37)	2.52 (1.34)	0.48	0.89	-0.35	0.14	–	–	–	–	–	–	–	–	–	–	-4.92	.00	.27
4	3.00 (1.21)	2.12 (1.08)	0.48	1.18	0.07	1.42	–	–	–	–	–	–	–	–	–	–	-13.94	.00	.69
7[i]	3.13 (1.25)	3.09 (1.25)	0.12	0.18	-0.47	-0.54	–	–	–	–	–	–	–	–	–	–	-0.64	.52	.03
10	2.25 (1.35)	1.89 (1.09)	0.96	1.43	0.18	2.02	–	–	–	–	–	–	–	–	–	–	-5.55	.00	.29
11	2.69 (1.43)	2.50 (1.37)	0.54	0.91	-0.55	0.17	–	–	–	–	–	–	–	–	–	–	-2.48	.00	.14
14	2.74 (1.29)	2.33 (1.21)	0.57	0.95	-0.12	0.58	–	–	–	–	–	–	–	–	–	–	-6.13	.00	.33
17	2.48 (1.33)	2.55 (1.41)	0.71	0.81	-0.19	-0.17	–	–	–	–	–	–	–	–	–	–	0.88	.37	.05
20	3.06 (1.33)	2.85 (1.53)	0.40	0.54	-0.38	-0.75	–	–	–	–	–	–	–	–	–	–	-2.85	.01	.15
21[i]	3.03 (1.23)	2.76 (1.24)	0.12	0.40	-0.54	-0.47	–	–	–	–	–	–	–	–	–	–	-4.06	.00	.22
29	2.88 (1.21)	2.35 (1.05)	0.57	0.99	0.05	1.32	–	–	–	–	–	–	–	–	–	–	-8.56	.00	.47
32	2.89 (1.25)	2.86 (1.40)	0.51	0.59	-0.14	-0.37	–	–	–	–	–	–	–	–	–	–	-0.38	.70	.02
34[i]	3.14 (1.18)	3.01 (1.25)	0.09	0.31	-0.47	-0.54	–	–	–	–	–	–	–	–	–	–	-2.06	.04	.11
Rigid																			
3[i]	3.54 (1.31)	4.01 (1.29)	-0.17	-0.38	-0.60	-0.53	–	-1.01	–	.03	–	1.11	–	1.15	–	.58	-6.64	.00	.36
6	3.05 (1.28)	2.54 (1.29)	0.41	0.86	-0.27	0.30	–	0.46	–	.03	–	1.05	–	1.04	–	.64	-7.35	.00	.40
8	3.10 (1.44)	2.41 (1.39)	0.30	0.94	-0.77	0.10	–	0.61	–	.03	–	1.23	–	1.15	–	.65	-8.99	.00	.49
13	3.14 (1.34)	2.62 (1.27)	0.34	0.65	-0.45	-0.05	–	0.36	–	.03	–	0.85	–	0.83	–	.69	-7.39	.00	.40
15[i]	3.14 (1.25)	3.01 (1.23)	0.03	0.17	-0.60	-0.56	–	-0.04	–	.03	–	0.91	–	0.91	–	.63	-1.90	.06	.10
19[i]	3.09 (1.37)	2.77 (1.38)	0.23	0.28	-0.79	-0.89	–	0.21	–	.03	–	1.17	–	1.15	–	.65	-4.37	.00	.23
22	3.05 (1.42)	2.97 (1.37)	0.37	0.55	-0.70	-0.40	–	0.00	–	.03	–	0.83	–	0.82	–	.73	-1.11	.27	.06

Table 1 (continued)

Item	M (SD)	M (SD)	Sk	K	Sk	K		Measure		SE		Infit		Outfit			t	p	d
24	3.34 (1.32)	2.82 (1.27)	0.13	0.55	-0.56	-0.20	–	0.15	–	.03	–	0.74	–	0.77	–	.72	-7.45	.00	.40
26	2.44 (1.29)	2.29 (1.16)	0.86	1.05	0.30	0.93	–	0.75	–	.04	–	1.25	–	1.26	–	.50	-2.20	.03	.12
30[i]	3.90 (1.23)	4.13 (1.14)	-0.53	-0.55	-0.00	0.21	–	-1.13	–	.03	–	0.83	–	0.85	–	.62	3.52	.00	.19
33	3.79 (1.21)	3.39 (1.35)	-0.00	0.16	-0.35	-0.63	–	-0.41	–	.03	–	1.21	–	1.24	–	.56	-5.85	.00	.31
35	3.29 (1.27)	2.92 (1.20)	0.25	0.54	-0.39	-0.01	–	0.05	–	.03	–	0.91	–	0.96	–	.60	-5.47	.00	.30

Note. Sk: skewness; K: kurtosis; Measure: location of each item along the continuum; SE: standard error; Infit/Outfit: mean squared residual. The significance level was fixed at < .001 using Bonferroni correction

[i] Inverted items

aleatory patterns of responses and short time lapses will be tracked and were reasons enough to remove collaboration reinforcers.

Data analysis

Descriptive statistics and independent sample t-tests for continuous variables were calculated to compare both BAPQ versions' means (statistical significance was set at $p < 0.001$ using Bonferroni correction). Effect sizes were estimated with Cohen's d using G*Power software v3.1 ($d = 0.2$ small effect size, 0.5 medium effect size, and 0.8 large effect size; Cohen, 1988). Dimensionality was checked using a principal components analysis of residuals (its interpretation differs from dimensionality studies using factor analysis). If unidimensionality was achieved, Rasch rating scale models (RSM[1]) for tests that have the same rating scales in all items were applied to each subscale separately (6-option Likert in the BAPQ; see Wright & Masters, 1982). RSM analyses were conducted in WINSTEPS software version 3.63.2 (Linacre, 2018) and plots were drawn with R software version 3.6.1 (R Core Team, 2019).

In the first place, item separation indices were tested to confirm that the sample size was adequate to estimate the item hierarchy. Indices over 3 assume that the person sample is large enough to verify the item severity hierarchy (Linacre, 2018) and to consider that it is highly likeable to find the same item placements if identical items are applied to another same-sized sample behaving similarly (Bond & Fox, 2015). In a second place, item fit (i.e., the degree to which the proposed item location can predict consistently participants' responses to these items) was analyzed (see Smith et al., 2008). Both, infit (sensitive to patterns of response of people with similar severity levels than one of the items) and outfit (sensitive to patterns of response of people with severity levels far from one of the items) mean-squared residual differences were estimated (values between 0.6 and 1.5 are interpreted as acceptable item fit; Lunz et al., 1990; Wright & Masters, 1982). In a third place, the locations of the items along each severity continuum were analyzed according to their content and the coherence of the item distribution was taken as a source of construct validity. In fourth place, the category probability curves of the items were analyzed. We explored whether the categories were ordered as expected and, whether they

---

[1] RSM assumes that the same rating scale (the thresholds) is shared among all items. It allows explore the essential properties of the BAPQ's subscales in six stages (corresponding with results subsections).

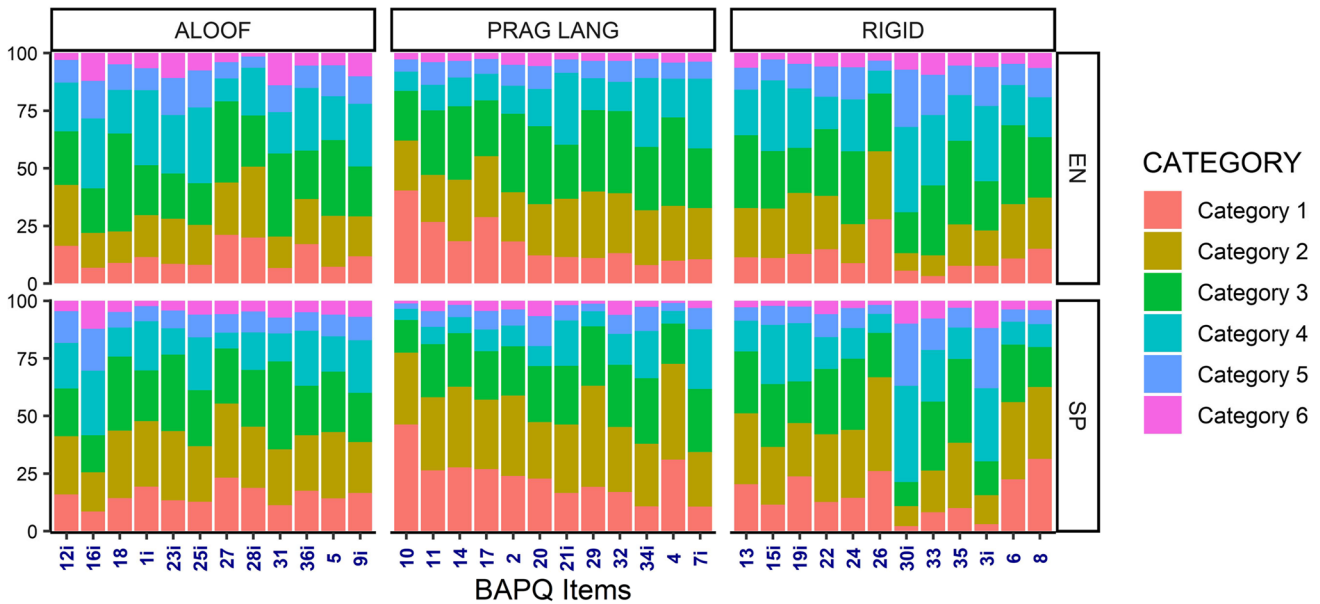# Percentage of Answers to each Response Category in Real Groups

Fig. 1 Item descriptive statistics: Percentage of responses to each response category in real groups

had the highest probability of adhesion, at least, at one point of the continuum. In fifth place, the efficacy of BAPQ to distinguish people along the different severity continuums was studied through people's separation indices and person reliability (i.e., the proportion of observed variance not due to measurement error; equivalent to the traditional test reliability). Person separation indices lower than 2 (person reliability < 0.80) indicate that the test cannot even distinguish between high and low severity levels (Bond & Fox, 2015). Furthermore, the precision of each set of items (i.e., Aloof, Rigid, and Pragmatic Language) throughout their corresponding severity continuum was examined using their test information functions (subscale in this case). Finally, we compared the performance between both versions of each subscale as depicted by their items; that is, DTF between the subscales of the English and the Spanish BAPQs with identity plots (for the whole analysis, consult Bond & Fox, 2015; see Fig. 5). These scatterplots represent item severity of each subscale version in each axis, an identity line symbolizing the expected values for severity invariance, and a pair of 95% quality control lines.

Results

The BAPQ-EN items showed higher averages and standard deviations than items in the BAPQ-SP (see Table 1 and Fig. 1). Particularly, nine, 10, and 11 items showed higher means in the Aloof, Pragmatic Language, and Rigid subscales

respectively. The means of the items in the BAPQ-EN as well as in the BAPQ-SP were low, some of them were between 3–4 and only two items in the BAPQ-SP were beyond 4.

Twenty-four items showed significant differences between both samples (significance was set at p < 0.001 using Bonferroni correction). Cohen's d revealed 22 items with small effect sizes (effect sizes; d = 0.2; items 1[i], 2, 3[i], 5, 6, 8, 9[i], 10, 11, 13, 14, 18, 19[i], 21[i], 24, 25[i], 28[i], 29, 30[i], 31, 33, 35), two mediums (d = 0.5; items 4 and 23[i]) and none item with large effect sizes (d = 0.8).

Subscale dimensionality and item separation indices

Results pointed to unidimensionality of both Aloof subscales (variances explained by measures were BAPQ-EN: 50.60% and BAPQ-SP: 55.00%). The BAPQ-EN Rigid subscale seemed to present lack of unidimensionality (variance explained by measure: 45.3%, eigenvalue = 2.769) with a secondary dimension (explained variance in the first contrast 12.60%, eigenvalue = 2.76, disattenuated correlation with de primary dimension = 0.41) formed by items 3[i], 15[i], 19[i], and 30[i]. The BAPQ-SP Rigid subscale was unidimensional (variance explained by measures: 51.10%). Finally, lack of unidimensionality was also observed in both Pragmatic Language subscales (explained variances by measures were BAQ-EN: 37.90%, BAPQ-SP: 42.00%). In the case of the BAPQ-EN, the secondary dimension (explained variance in first contrast 14.2%, eigenvalues = 2.74, disattenuated correlation with the primary dimension = 0.21)

was formed by items 7^i, 21^i, and 34^i. In BAPQ-SP, secondary dimension (explained variance in first contrast 9.50%, eigenvalue = 1.97, disattenuated correlation with the primary dimension = 0.49) was formed by items 14, 17, 20, and 32. Lack of unidimensionality impeded to perform further RSM analysis on both Pragmatic Language subscales and Rigid of the BAPQ-EN.

Regarding item separation indices, for the BAPQ-EN, they were between [7.23, 7.64] for Aloof items. For the BAPQ-SP, they were between [7.98, 8.37] for Aloof items and [16.23, 16.94] for Rigid ones.

Correlations of the persons' IRT scores between subscales in BAPQ-SP was: Aloof-Rigidity 0.63, p < 0.001.

Item fit

Infit and outfit mean-squared residual differences for Aloof and Rigid subscales are displayed in Table 1. Regarding the BAPQ-EN Aloof subscale, only Item 18 was slightly upon (1.65) the established boundaries on its outfit index (i.e., [0.6, 1.5]). None item in the BAPQ-SP Aloof and Rigid subscales showed a misfit. Item-total correlations showed adequate values for the BAPQ-EN and BAPQ-SP. The correction of the mistranslation in item 23 resulted in good fit indices and a positive correlation between the item and the subscale (see Table 1) pointing to its adequate functioning similar to what was observed for its English version (BAPQ-EN Aloof subscale).

Item locations

The estimated item and person locations are exhibited in Fig. 2 while item locations and estimated parameters are presented in Table 1.

Overall, Aloof item locations showed higher levels of severity (about one item standard deviation above the item mean) than the average Aloof level in the samples for both BAPQ versions. Additionally, some items shared similar locations in the continuum although they refer to different contents (e.g., items 5 and 18 in the BAPQ-EN or items $5, 12^i$, 31, and 36^i in the BAPQ-SP). The item severity mean in the Rigid subscale in the BAPQ-SP was about one item standard deviation of the mean of the items. Items 15^i, 22, and 35; items 6 and 13; and items 19^i and 24 also clustered in the continuum. The remaining five items were spread in different positions below and above the item severity mean.

Category probability curves

Aloof subscale in the BAPQ-EN and Aloof and Rigid in the BAPQ-SP showed adequate category functioning (see Fig. 3).

People separation indices and person reliability

The separation index was between [2.60, 3.08] for the BAPQ-EN Aloof subscale ([2.81, 3.29] in the BAPQ-SP), and the scale reliability was between [0.87, 0.90] ([0.89, 0.92] in the BAPQ-SP) that means that both Aloof subscales were able to distinguish almost among three statistically different strata. In the BAPQ-SP Rigid subscale, the separation index was between [2.37, 2.70] and the scale reliability was between [0.85, 0.88], which means that it was able to statistically distinguish between two different strata.

Subscale information functions, as well as standard errors of measurements, can be observed in Fig. 4. The point where the Aloof subscale reached its highest precision level corresponded, approximately, with the raw scores 13.14 in BAPQ-EN and 13.22 in BAPQ-SP. The point where Rigid in BAPQ-SP reached its highest precision level corresponded, approximately, with the 13.39 raw score (Fig. 5).

Differential Test Functioning (DTF)

Finally, we checked the Aloof subscale DTF between the English and the Spanish BAPQs (see Fig. 5).

Ten out of twelve items of the Aloof subscale showed DIF. Only items 5 and 9^i had similar functioning. With similar levels of Aloofness, items 1^i, 18, 23^i, and 25^i would be less likely to be endorsed by the Spanish sample than by the English sample, while the opposite happened with items 12^i, 16^i, 27, 28^i, 31, and 36^i.

Discussion

In this study, we applied IRT to further study two BAPQ versions (BAPQ-EN and BAPQ-SP; Hambleton et al., 1991). Additionally, we tested the stability of the items' hierarchy and which BAPQ item severities function differentially across versions (Hambleton, 2006).

Significant differences were observed when the same items were applied to both samples (the larger ones are explained further down) being the BAPQ-EN sample the one with overall higher item means. However, only small and medium effect sizes indicated no critical relevance.

Unidimensionality problems in both Pragmatic Language subscales and Rigid one in the BAPQ-EN impeded further rating scale analysis (Bond & Fox, 2015) and, congruently with previous findings, provides negative validity evidence regarding subscales' internal structures. Particularly, Pragmatic Language adverse evidence was reported in Godoy-Giménez et al. (2018), Sasson et al. (2013a), and Sharma
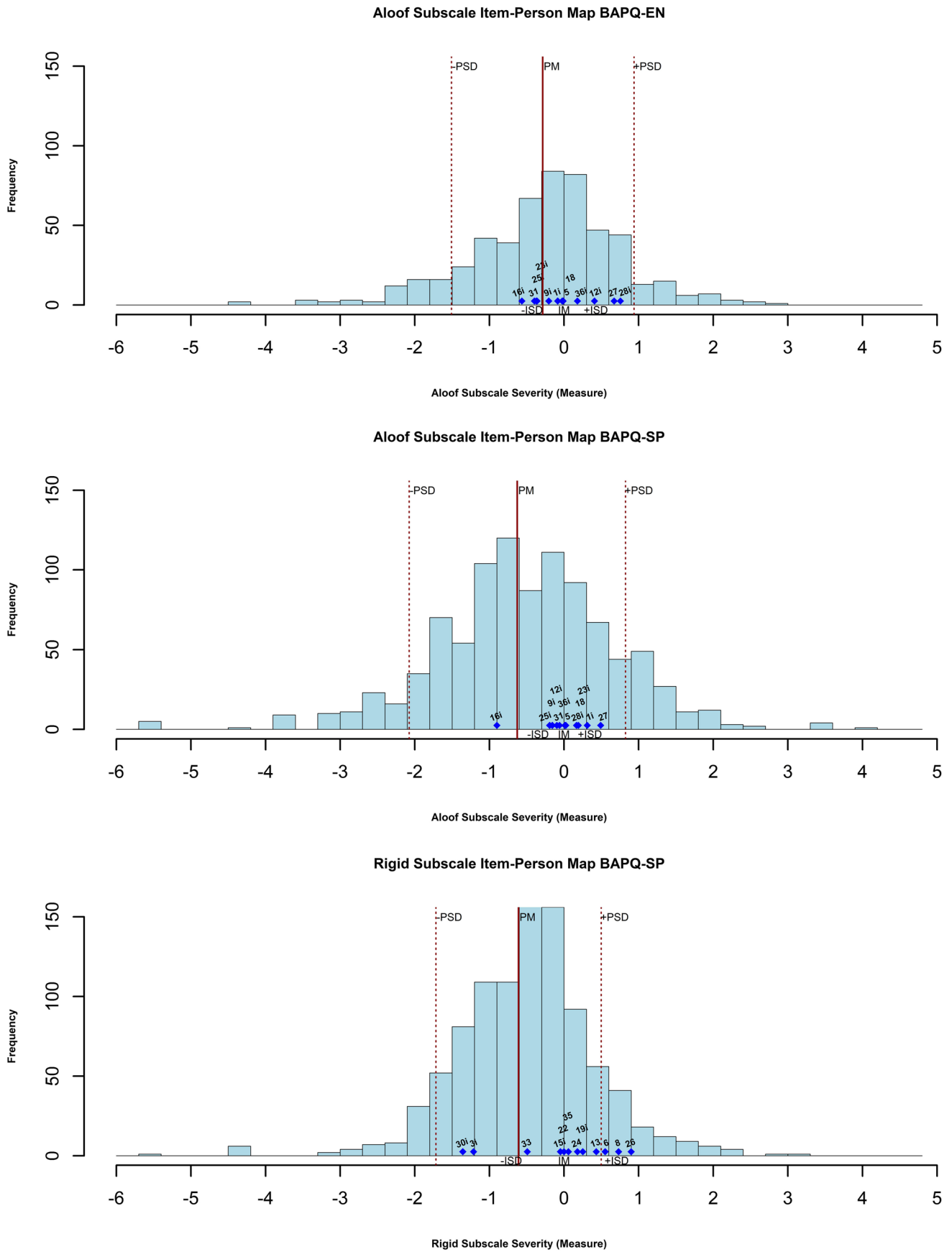
Fig. 2 Person-item maps of the Aloof BAPQ-EN and Aloof and Rigid subscales BAPQ-SP

and Bhushan (2018); Rigid adverse evidence was reported in Lin et al. (2021), and adverse evidence of both subscales in Stojković et al. (2018). These findings could be better understood under the light of the most updated BAP operationalization (Godoy-Giménez et al., 2021) which revolves around two core domains paralleling it with the last ASD definition (APA, 2013) and includes some variations in the test content.

Regarding item fit, in the BAPQ-EN, we consider that the item in the Aloof subscale (Item 18) that exhibited little misfit could be explained by a lack of concreteness in the item content (any person rating high or low in BAP could agree with it because being polite is independent of the preference for social interactions), yet this slight item misfit could not degrade the measure (Linacre, 2018). By contrast, in the BAPQ-SP, this item did not show any misfit. As this item is located in different parts of the severity continuums of Aloof in both samples, we could suggest that cultural aspects could be playing a role in how different cultures interpret the item. In this regard, if we consider that the item is characterized by a lack of concreteness, it could be possible that the Spanish group understands the item more concretely. However, it would be also reasonable to think that some cultures give more importance to politeness than others. Finally, it is worth mentioning that this item, as other items of the BAPQ, is linked to particularity as in the BAPQ some items are
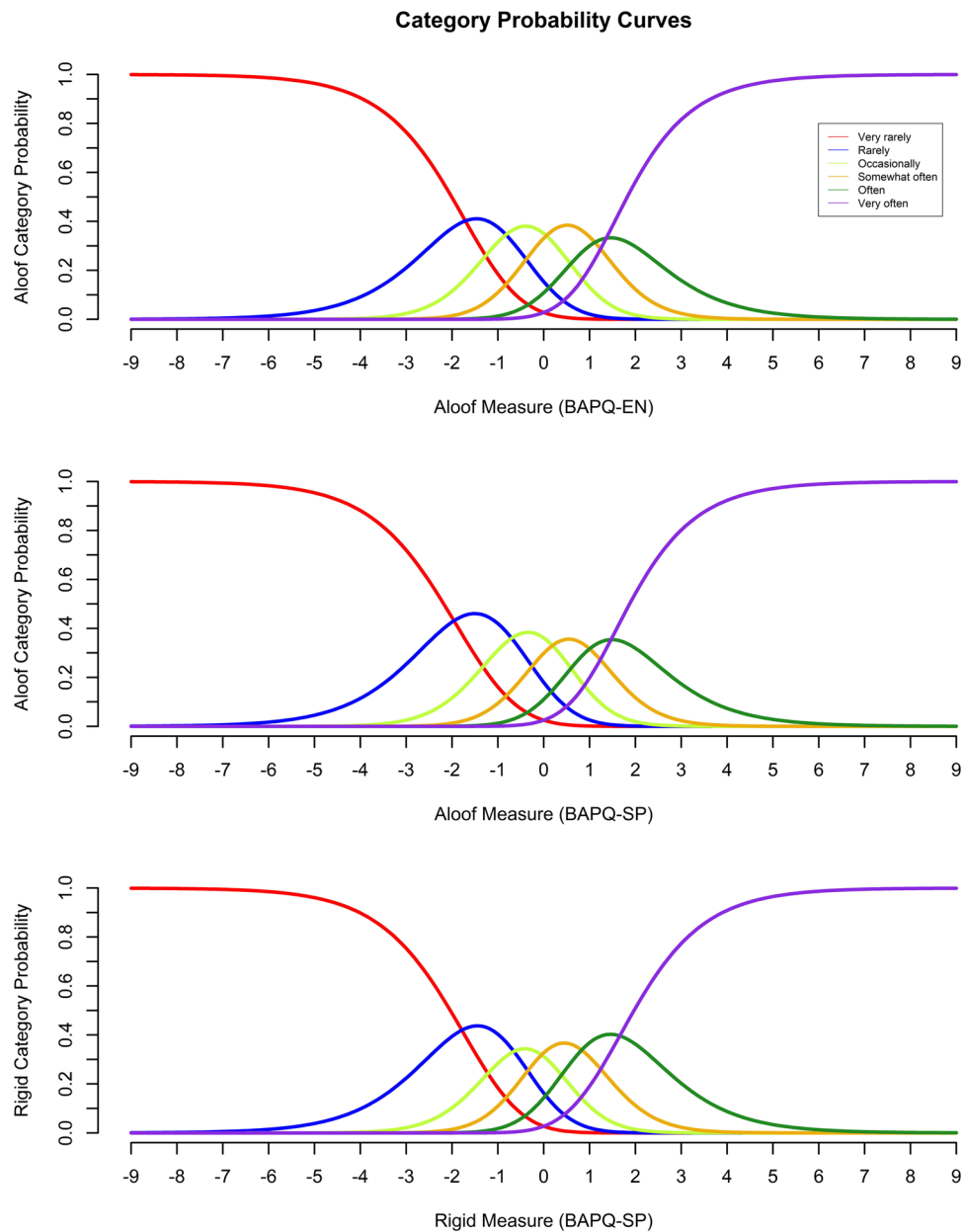
Fig. 3 Category probability curves of the Aloof BAPQ-EN and Aloof and Rigid subscales BAPQ-SP

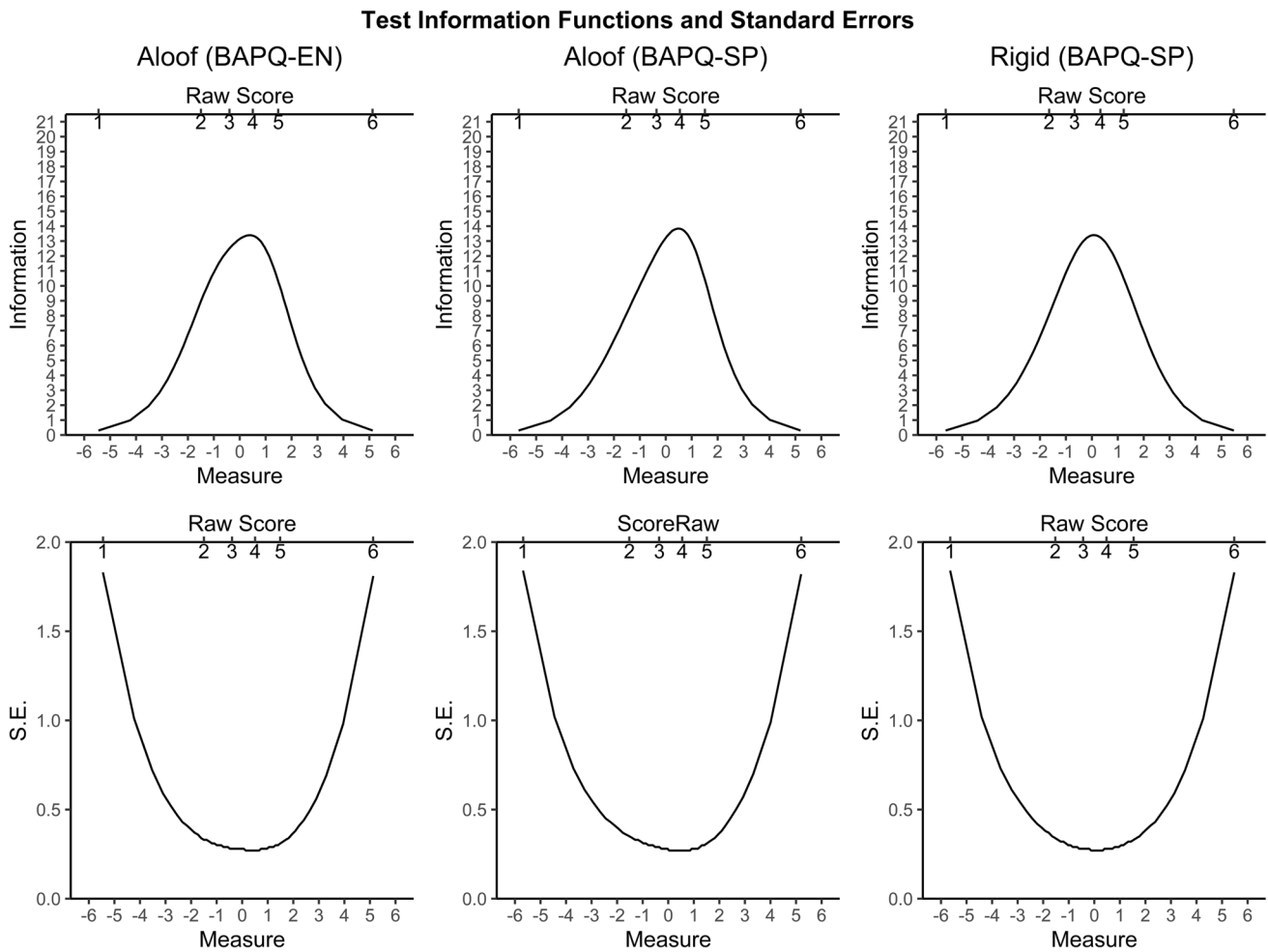## Test Information Functions and Standard Errors

Fig. 4 Test information and standard error functions of the Aloof BAPQ-EN and Aloof and Rigid subscales BAPQ-SP. Note. T. I. F. = test information function; S.E. = Standard Error

specifically referred to as casual interactions with acquaintances (see Godoy-Giménez et al., 2018; Hurley et al., 2007). Consequently, it could be possible that both samples have differed in how they applied this particularity at the moment to respond to the item. Any other item has not shown misfit nor item 23 from the Aloof subscale in the BAPQ-SP. This result is relevant since, together with a good item-total scale correlation, it supports the alternative translation of this item. In sum, we can conclude that the results of the present study have reinforced the alternative translations of the Spanish items 4 and 23 (item-total correlation values of Spanish item 4 [0.64] and 23 [0.57] spoke in favour of this). Thirdly, average item severities for both Aloof and BAPQ-SP Rigid subscales, were upper than the average severity level of both community samples. Consequently, subscales' reliability reaches its higher value always for upper scores than both samples mean. This confirms our hypotheses that the items will mainly be situated in middle-upper levels of the continuum indicating that the BAPQ severity levels are more targeted for assessing high levels of BAP, like those

observed in parents of ASD children as originally intended in Hurley et al., 2007. Thus, for those studies focused on community or severity-diverse populations (e.g., Faso et al., 2016; Morrison et al., 2018), it could be advisable to use tests with subtler BAP indicators to enhance the measurement accuracy at lower BAP severity levels.

Regardless of this issue, the items and persons have congruently been scaled according to their severity in both Aloof subscales and the Spanish Rigid one and that implied that the items of the BAPQ subscales are susceptible to being scaled along a severity dimension (severity hierarchies will also be discussed below in terms of validity evidence). Furthermore, our analyses revealed adequate category functioning for both Aloof subscales and the Rigid BAPQ-SP. Thus, even if the test counts on a 6-point scale, we have reported that all the rating options functioned and are most probably chosen at some point in the continuum in both samples. In the same line, adequate reliability and separation indices of both Aloof and the Spanish Rigid subscales inform that subscale hierarchy can statistically differentiate at least among
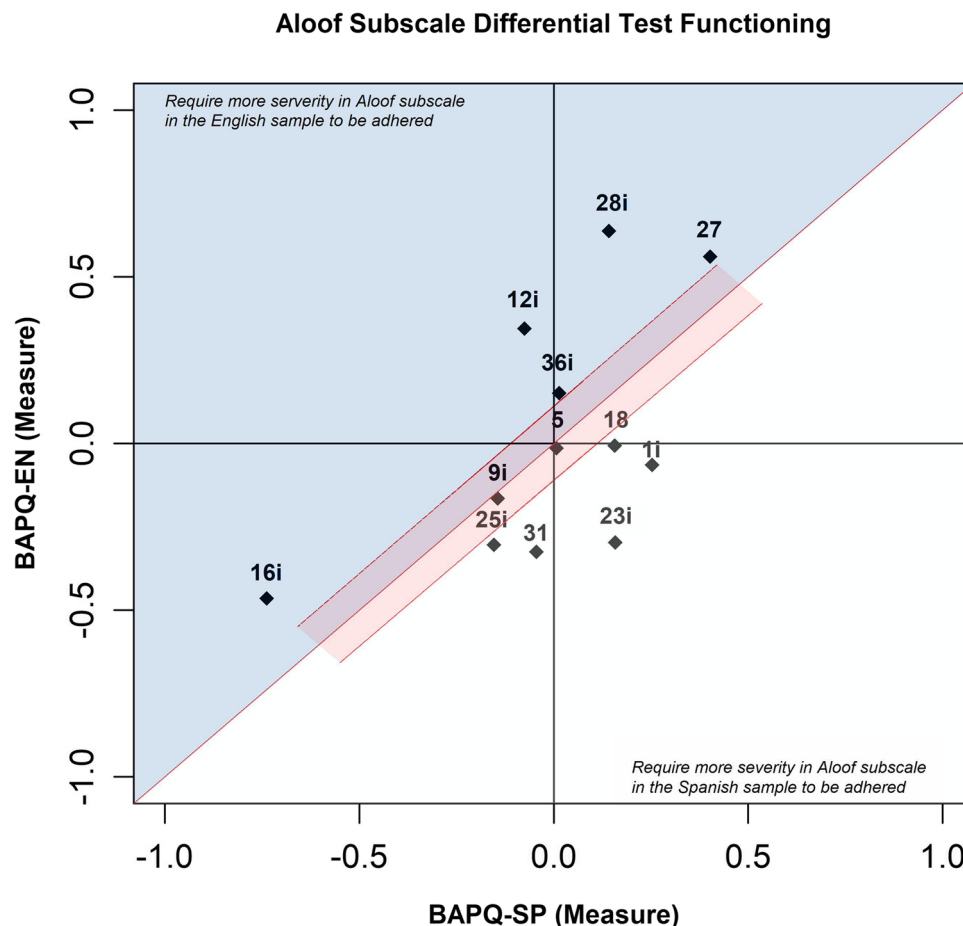
Aloof Subscale Differential Test Functioning

Fig. 5 Identity plot with each Aloof subscale**.** Note: Scatterplots represent the severity location of the items scores regarding each BAPQ version (BAPQ-EN and BAPQ-SP). Points placed on the x-axis represent the respective item values along the BAPQ-EN severity continuum. Points placed on the y-axis represent the respective item values along the BAPQ-SP severity continuum

two different strata of BAP severity (i.e., high and low BAP). This represents an important highlight since it goes in line with the ASD severity dimensional approach and opens the door to locate the BAP inside the autism continuum. In this regard, BAP behaviours would not define a qualitatively different group but rather it would comprehend people who could express autistic-like behaviours, at least, at higher or milder severity that makes them more or less functionally independent.

Our second objective was to compare the differential functioning of both Aloof subscales between the Spanish adaptation of the BAPQ (Godoy-Giménez et al., 2018) and its original version (Hurley et al., 2007). Two items showed invariant severity and thus, they would be equally endorsed in both samples by persons with the same severity level of aloofness. By contrast, the lack of invariance in the rest of the items hindered the comparison of the severity hierarchies of both Aloof subscales between the English and Spanish versions.

On the other hand, the location and content of those invariant items need to be discussed. As such, we consider that a mild introverted behaviour could be not to enjoy being in social situations (item 9[i]: "I enjoy being in social situations") which does not imply that the respondent avoids spending time with a few close friends or relatives. By contrast, expressing an instrumental use of the acts of socializing, in general (also involves interactions with relatives, close friends, and partners) could be taken as a more severe indicator of aloofness (item 5: "I would rather talk to people to get information than to socialize"). Nevertheless, it is worth mentioning that this last behaviour still locates the person in social interaction and surrounded by others. Coherently, authors have found that ASD parents have shown less interest in purely social interactions, and report having fewer and lower quality friendships (Faso et al., 2016).

Even though the rest of the items have not resulted invariant, both Aloof hierarchies share similar severity order in both groups. This also deserves to be highlighted as it could

be taken as additional validity evidence. Specifically, six items of both Aloof hierarchies have shown the same order (from upper to lower locations in the continuum: item 27, 18, 5, 9[i], 25[i], 16[i]). Items that are sharing near placements in both Aloof hierarchies are going to be commented together (upper locations denote more aloofness-like behaviours or more Aloof severity). Item 27 ("Conversation bores me") is located in upper locations in both severity continuums and it refers to an open expression of disinterest for establishing social communication in casual interactions with acquaintances. Items expressing less amount of Aloof were 18 ("When I make conversation it is just to be polite"), 5 ("I would rather talk to people to get information than to socialize"), and 9[i] ("I enjoy being in social situations") which also express this disinterest for social interactions but more subtly and, even when people do not give chance for establishing longer interactions (and do it only for being polite), they do participate in social communication. Finally, items 25[i] and 16[i] are below the mentioned items. Coherently with severity scaling of the items, those two items refer to not having an active pursuit of social situations, social involvement, or social enjoyment, but still, it does not imply being in social situations and participating in them. We could consider that they express milder severity behaviours (e.g., item 25[i] "I feel like I am really connecting with other people"; item 16[i] "I look forward to situations where I can meet new people"). Importantly, this order in both Aloof hierarchies is also congruent with ASD specifiers in the DSM-V (APA; 2013): Social impaired behaviours at ASD levels could also vary from a decreased interest in social interactions (severity level 1, individual are more functional), to limited initiation of social interactions (severity level 2), and to unusual social approaches restricted to meet needs (severity level 3, individual are less functional; ASD severity specifiers; APA, 2013). This is also an important highlight because both shared Aloof hierarchy orders would be susceptible of being reinserted in the autistic severity continuum (De Groot & Van Strien, 2017; Hoekstra et al., 2008) and also provide validity evidence about the fact that the BAP does not only shares the same autistic behaviours but also that the scalability of the severity of those behaviours follows the same path than in more severe levels of ASD behaviours (APA, 2013). By contrast, some items were located in a different order in both Aloof hierarchies. For example, items 1[i] and 23[i] are situated below item 27 in upper locations in the BAPQ-SP and item 31[i] in milder locations between items 5 and 9[i] while these items are all located below item 5 in the Aloof BAPQ-EN. In the case of the BAPQ-EN, items 28[i], 12[i], and 36[i] are surrounding item 27 (upper locations of the severity continuum) while in the Aloof BAPQ-SP those items are targeting milder placements. Items with different placements may be referring to varying social situations. In the case of Aloof BAPQ-SP, severe item 1[1] ("I like being around other

people") and 23[i] ("I am good at making small talk") refers to social enjoyment and involvement but with the nuance of being surrounded by many people or known people. In the BAPQ-EN, upper items 28[i], 12[i], and 36[i] express being warm and friendly in social interactions (28[i] "I am warm and friendly in my interactions with others"; 12[i] "People find it easy to approach me"; 36[i] "I enjoy chatting with people"). For the item severities that did not remain stable between versions, Hambleton (1994) pointed to "genuine cultural specifics" as a possible explanation of item variance. In this regard, and linked to what we have explained earlier about item 18, we could speculate if this could be suggesting that being polite is an important aspect of social interactions for the English sample while social involvement could be more important in the case of the Spanish one. Nevertheless, these hypotheses should be studied in the future. In the same way, the hierarchical three-level framework that explains how culture affects the perception and diagnosis of psychiatric disorders proposed by Rogler (1993) could be applied to ASD and, in extension, to BAP levels of severity. Accordingly, cultural norms mediate both the endorsement of symptoms and how people rate the severity of symptoms. Therefore, two participants with the same BAP severity but from different countries could express different behaviours responding differently to the BAPQ.

Since the severities of some Aloof items are sensible to the influence of diverse cultural aspects, the direction, as well as the magnitude of these influences, should be established if authors want to further study the interactions among ASD and BAP expressions and the culture. This implies integrating a comprehensive approach taking into account both the invariant and the non-invariant item severities and hierarchies, paying special attention to the influence of the cultures on the development of BAP-related behaviours and preferences but also on their expression either in a natural context or in a test. Ethnically-based cultural norms would modulate the perception of socially undesirable mental symptoms, and, ultimately, the own endorsement of those symptoms (Matson et al., 2017). As such, how people report the behaviours they think they are expressing and their severity may be influenced by the aspects that are considered more worrying within the culture. For example, some studies carried out in the United States have suggested that American parents tend to be more concerned about language delays (Coonrod & Stone, 2004) while Indian parents would tend to have early concerns about social difficulties (Daley, 2004), and Latina mothers may be worried about their child temperament (Ratto et al., 2016).

On the other hand, the severity hierarchy of the Rigid BAPQ-SP should be equally commented. Items sharing near placements of the Rigid continuum are going to be discussed together (upper locations denote more rigidity or more Rigid severity). Items 26 ("People get frustrated by my unwillingness to bend") and 8 ("I have to warm myself up to the idea of visiting an unfamiliar place") which refers to strong stubborn-

ness or an extreme need for sameness and daily routines were situated above the rest of the items. Below them, we can also find items 13 and 6 that also express a preference for sameness and routines (e.g., 13 "I feel a strong need for sameness from day to day", 6 "People have to talk me into trying something new") and 24 ("I act very set in my ways") and 19[i] (" I look forward to trying new things") which also reflect stubbornness, sameness, and routines but more subtly without the nuances of "warm myself" or "unwillingness to bend" that comprehend more severe items (i.e., 26 and 8). Further down in the severity continuum there are items concerning struggling with alternative work procedures (generally about how things must be done) or changes in daily routines (e.g., 35 "I keep doing things the way I know, even if another way might be better", 22 "I have a hard time dealing with changes in my routine" and 15[i] "I am flexible about how things should be done") but, importantly, items do not specify that those difficulties impede to conduct alternative protocols or alternative plans. In lower locations in the continuum, item 33 refers to a preference for fixed work protocol ("I like to closely follow a routine while working") but does not express the difficulty component. Finally, following severity order, last placements are taken by items that express a preference for daily routines or not dealing with unexpected plan changes (3[i] "I am comfortable with unexpected changes in plans"; 30[i] "I alter my daily routine by trying something different") which we consider behaviours to which almost every person would be comfortable to adhere to. Additionally, we should appreciate that the resulting order of Rigid hierarchy is aligned with the ASD.

Like in the case of Aloof subscales, the resulting order of the Rigid BAPQ-SP hierarchy could open the door to insert behaviours at BAP levels of severity inside the autism continuum (De Groot & Van Strien, 2017; Hoekstra et al., 2008). In this sense, stubbornness, persisting routines and a strong need for sameness (i.e., Item 26) could be nearer clinical levels of severity, mostly when inflexible behaviours cause significant interference with individual functioning in one or more contexts (Item 15[i]; level 1), it comprises difficulty coping with change (Item 8; level 2), and those behaviours marking interfering in all the spheres (level 3; ASD severity specifiers; APA, 2013). Directly linked with this, some autistic-behaviours found upper in the Rigid BAP-SP severity hierarchy are similar to those observed in Obsessive–Compulsive Personality Disorder (e.g., overly rigid and/or stubborn, perfectionism, and very strict work standards; APA, 2013) or even to obsessive–compulsive disorder ones (e.g., obsessional thoughts; APA, 2013). Of importance, obsessive–compulsive behaviours have been reported comorbid to autism spectrum disorder (e.g., Meier et al., 2015; Micali et al., 2004) and this would also support the resulting Rigid severity order and the reinsertion of this dimension inside the autism continuum.

Altogether, the presented results lead us to think that the BAPQ could have been a useful measurement tool for meas-

uring the BAP in English populations with middle-high BAP severity levels according to the original BAP operationalization (Piven et al., 1997a). Nevertheless, this study, in conjunction with others, has provided adverse evidence about the internal structure of the BAPQ (e.g., Godoy-Giménez et al., 2018; Lin et al., 2021; Sasson et al., 2013a; Sharma and Bhushan, 2018; Stojković et al., 2018). These findings could be better understood considering the most updated BAP operationalization (Godoy-Giménez et al., 2021; Morrison et al., 2018; Sasson et al., 2013b) and suggest that a new BAP test based on an updated BAP definition (as the middle expression of the current two-dimensions operationalization of the ASD; APA, 2013; see Godoy-Giménez et al., 2021) should be built. The new test would aim to measure BAP covering all autism subthreshold levels of expression including those in the general population. Additionally, to accomplish the requirements of the current international research context we believe that this new test should aim to inform about how the severity of the items could vary among cultures.

Declarations

Compliance with Ethical Standards This study received ethics approval from the local Human Research Ethics Committees of both universities. The recruitment process was conducted under the approved guidelines and the Declaration of Helsinki. Written informed consent was obtained from all participants in advance. Participants in both samples received two-course credits for their collaboration.

References

Álvarez-Couto, M., García-Villamisar, D., Sáez-Suanes, G. P., & Roquete, M. D. O. (2021). Variables predicting the association between autistic traits and externalizing symptoms among young adults. Advances in Autism. Vol. ahead-of-print No. ahead-of-print. https://doi.org/10.1108/AIA-06-2020-0036

American Psychiatric Association. (1994). Diagnostic and statistical manual of mental disorders (DSM-IV®). American Psychiatric Pub.

American Psychiatric Association. (2013). Diagnostic and statistical manual of mental disorders (5th ed.). https://doi.org/10.1176/appi.books.9780890425596

Azizan, N. H., Mahmud, Z., Rambli, A.(2020). Rasch Rating Scale Item Estimates using Maximum Likelihood Approach: Effects of Sample Size on the Accuracy and Bias of the Estimates. International Journal of Advanced Science and Technology Vol. 29, No. 4s, pp. 2526 - 2531.

Bang, P., Strömberg, M., Meera, S. S., & Igelstrom, K. (2021). Brief Report: The Broad Autism Phenotype in Swedish Parents of Children with and without Autism Spectrum Conditions. Journal of Autism and Developmental Disorders. 1-8. https://doi.org/10.1007/s10803-021-05302-3

Berry, J. W., Poortinga, Y. H., Segall, H. H., & Dasen, P. R. (1992). Cross-cultural psychology: Research and applications. Cambridge University Press.

Bölte, S., Poustka, F., & Constantino, J. N. (2008). Assessing autistic traits: Cross-cultural validation of the social responsiveness scale (SRS). Autism Research, 1(6), 354–363. https://doi.org/10.1002/aur.49

Bond, T. G., & Fox, C. M. (2015). Applying the Rasch model: Fundamental measurement in the human sciences (3rd ed.). L Erlbaum.

Broderick, N., Wade, J. L., Meyer, J. P., Hull, M., & Reeve, R. E. (2015). Model Invariance Across Genders of the Broad Autism Phenotype Questionnaire. Journal of Autism and Developmental Disorders, 45, 3133–3147. https://doi.org/10.1007/s10803-015-2472-z

Camodeca, A., & Voelker, S. (2016). Automatic and controlled processing and the Broad Autism Phenotype. Psychiatry Research, 235, 169–176. https://doi.org/10.1016/j.psychres.2015.11.012

Cohen, J. 1988. Statistical Power Analysis for the Behavioral Sciences, 2nd Edition. Routledge

Coonrod, E. E., & Stone, W. L. (2004). Early concerns of parents of children with autistic and nonautistic disorders. Infants & Young Children, 17(3), 258-268.

Daley, T. C. (2004). From symptom recognition to diagnosis: children with autism in urban India. Social science & medicine, 58(7), 1323-1335. https://doi.org/10.1016/S0277-9536(03)00330-7

De Groot, K., & Van Strien, J. W. (2017). Evidence for a broad autism phenotype. Advances in Neurodevelopmental Disorders, 1(3), 129–140. https://doi.org/10.1007/s41252-017-0021-9

Faso, D. J., Corretti, C. A., Ackerman, R. A., & Sasson, N. J. (2016). The broad autism phenotype predicts relationship outcomes in newly formed college roommates. Autism, 20(4), 412–424. https://doi.org/10.1177/1362361315585733

Godoy-Giménez, M., González-Rodríguez, A., Cañadas, F., Estévez, A. F., & Sayans-Jiménez, P. (2018). Psychometric Properties of the Spanish Version of the Broad Autism Phenotype Questionnaire: Strengths, Weaknesses, and Future Improvements. Journal of Autism and Developmental Disorders, 48(3), 770–778. https://doi.org/10.1007/s10803-017-3438-0

Godoy-Giménez, M., González-Rodríguez, A., Cañadas, F., Estévez, A. F., & Sayans-Jiménez, P. (2021). Is it possible to assess the two-domain definition of the Broad Autism Phenotype using the available measurement tools? Journal of Autism and Developmental Disorders. https://doi.org/10.1007/s10803-021-05158-7

Hambleton, R. K. (1994). Guidelines for adapting educational and psychological tests: A progress report. European Journal of Psychological Assessment, 10(3), 229–244.

Hambleton, R. K. (2006). Good Practices for Identifying Differential Item Functioning. Medical Care, 44(11), S182-S188. Retrieved from http://www.jstor.org/stable/41219518

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). Fundamentals of item response theory. Sage Publications.

Hoekstra, R. A., Bartels, M., Cath, D. C., & Boomsma, D. I. (2008). Factor structure, reliability and criterion validity of the Autism-Spectrum Quotient (AQ): A study in Dutch population and patient groups. Journal of Autism and Developmental Disorders, 38(8), 1555–1566. https://doi.org/10.1007/s10803-008-0538-x

Hofstede, G. (2001). Culture's consequences: Comparing values, behaviours, institutions, and organizations across nations. Sage publications.

Hurley, R. S., Losh, M., Parlier, M., Reznick, J. S., & Piven, J. (2007). The broad autism phenotype questionnaire. Journal of Autism and Developmental Disorders, 37, 1679–1690. https://doi.org/10.1007/s10803-006-0299-3

Ingersoll, B., & Wainer, A. (2014). The Broader Autism Phenotype. In F. R. Volkmar, R. Paul, S. J. Rogers, & K. A. Pelphrey (Eds.), Handbook of Autism and Pervasive Developmental Disorders (4th ed.). John Wiley & Sons. https://doi.org/10.1002/9781118911389

Jakobson, L. S., Pearson, P. M., Kozub, Z., Hare, C., & Rigby, S. N. (2018). Links between traits associated with the broad autism phenotype and empathy and young adults' ability to decode speaker intentionality. Research in Autism Spectrum Disorders, 50, 11–21. https://doi.org/10.1016/j.rasd.2018.03.001

Jamil, R., Gragg, M. N., & DePape, A.-M. (2017). The broad autism phenotype: Implications for empathy and friendships in emerging adults. Personality and Individual Differences, 111, 199–204. https://doi.org/10.1016/j.paid.2017.02.020

Kim, H. U. (2012). Autism across cultures: Rethinking autism. Disability & Society, 27(4), 535–545. https://doi.org/10.1080/09687599.2012.659463

Landa, R., Piven, J., Wzorek, M., Gayle, J., Chase, G., & Folstein, S. (1992). Social language use in parents of autistic individuals. Psychological Medicine, 22, 245–254. https://doi.org/10.1017/S0033291700032918

Landry, O., & Chouinard, P. A. (2016). Why we should study the broader autism phenotype in typically developing populations. Journal of Cognition and Development, 17, 584–595. https://doi.org/10.1080/15248372.2016.1200046

Lin, J., Fu, X., Dai, S., Zeng, Y., Yang, R., Liu, R., ... & Ou, J. (2021). The Broad Autism Phenotype Questionnaire (BAPQ): Strengths, weaknesses, and future improvements in the Chinese version. Asian Journal of Psychiatry, 60, 102655. https://doi.org/10.1016/j.ajp.2021.102655

Linacre J. M. (2002). Optimizing rating scale category effectiveness. Journal of applied measurement, 3(1), 85–106.

Linacre, J. M. (2018). Winsteps® Rasch measurement computer program. User's Guide. Beaverton, or: Winsteps.com.

Lunz, M. E., Wright, B. D., & Linacre, J. M. (1990). Measuring the impact of judge severity on examination scores. Applied measurement in education, 3(4), 331-345. https://doi.org/10.1207/s15324818ame0304_3

Masters, G. N., & Wright, B. D. (1984). The essential process in a family of measurement models. Psychometrika, 49(4), 529–544. https://doi.org/10.1007/bf02302590

Matson, J. L., Matheis, M., Burns, C. O., Esposito, G., Venuti, P., Pisula, E., Misiak, A., Kalyva, E., Tsakiris, V., Kamio, Y., Ishitobi, M., & Goldin, R. L. (2017). Examining cross-cultural differences in autism spectrum disorder: A multinational comparison from Greece, Italy, Japan, Poland, and the United States. European Psychiatry: The Journal of the Association of European Psychiatrists, 42, 70–76. https://doi.org/10.1016/j.eurpsy.2016.10.007

Meera, S. S., Girimaji, S. C., Seshadri, S. P., Philip, M., Shivashankar, N., Morgan, P., & Piven, J. (2015). Translation of the Broad Autism Phenotype Questionnaire to an Indian language: A

description of the process. *Asian Journal of Psychiatry, 15*, 62–67. https://doi.org/10.1016/j.ajp.2015.04.013

Meier, S. M., Petersen, L., Schendel, D. E., Mattheisen, M., Mortensen, P. B., & Mors, O. (2015). Obsessive-Compulsive Disorder and Autism Spectrum Disorders: Longitudinal and Offspring Risk. *PLoS ONE, 10*(11), e0141703. https://doi.org/10.1371/journal.pone.0141703

Micali, N., Chakrabarti, S., & Fombonne, E. (2004). The Broad Autism Phenotype: Findings from an Epidemiological Survey. *Autism, 8*, 21–27. https://doi.org/10.1177/1362361304040636

Mitchell, M., & Jolley, J. (2013). Research design explained (8th ed.). Belmont, CA:Wadsworth.

Morrison, K. E., Chambers, L. K., Faso, D. J., & Sasson, N. J. (2018). The content and function of interests in the broad autism phenotype. *Research in Autism Spectrum Disorders, 49*, 25–33. https://doi.org/10.1016/j.rasd.2018.02.002

Neil W. Scott; Peter M. Fayers; Neil K. Aaronson; Andrew Bottomley; Alexander de Graeff; Mogens Groenvold; Chad Gundy; Michael Koller; Morten A. Petersen; Mirjam A.G. Sprangers (2009). A simulation study provided sample size guidance for differential item functioning (DIF) studies using short scales. 62(3), 0–295. https://doi.org/10.1016/j.jclinepi.2008.06.003

Piven, J., Palmer, P., Jacobi, D., Childress, D., & Arndt, S. (1997a). Broader autism phenotype: Evidence from a family history study of multiple-incidence autism families. *American Journal of Psychiatry, 154*(2), 185–190. https://doi.org/10.1176/ajp.154.2.185

Piven, J., Palmer, P., Landa, R., Santangelo, S., Jacobi, D., & Childress, D. (1997b). Personality and language characteristics in parents from multiple-incidence autism families. *American Journal of Medical Genetics, 74*(4), 398–411. https://doi.org/10.1002/(SICI)1096-8628(19970725)74:4%3c398::AID-AJMG11%3e3.0.CO;2-D

Ratto, A. B., Reznick, J. S., & Turner-Brown, L. (2016). Cultural effects on the diagnosis of autism spectrum disorder among Latinos. *Focus on Autism and Other Developmental Disabilities, 31*(4), 275–283. https://doi.org/10.1177/1088357615587501.

R Core Team (2019). R: A Language and Environment for Statistical Computing. Vienna: R Foundation for Statistical Computing. Available online at: https://www.R-project.org/

Rogler, L. H. (1993). Culturally sensitizing psychiatric diagnosis: A framework for research. *Journal of Nervous and Mental Disease, 181*(7), 401–408. https://doi.org/10.1097/00005053-199307000-00001

Sasson, N. J., Faso, D. J., Parlier, M., Daniels, J. L., & Piven, J. (2014). When father doesn't know best: Selective disagreement between Self-Report and informant report of the broad autism phenotype in parents of a child with autism. *Autism Research, 7*(6), 731–739. https://doi.org/10.1002/aur.1425

Sasson, N. J., Lam, K. S. L., Childress, D., Parlier, M., Daniels, J. L., & Piven, J. (2013a). The Broad Autism Phenotype Questionnaire: Prevalence and Diagnostic Classification. *Autism Research, 6*, 134–143. https://doi.org/10.1002/aur.1272

Sasson, N. J., Lam, K. S., Parlier, M., Daniels, J. L., & Piven, J. (2013b). Autism and the broad autism phenotype: Familial patterns and intergenerational transmission. *Journal of Neurodevelopmental Disorders, 5*, 11. https://doi.org/10.1186/1866-1955-5-11

Scott, N. W., Fayers, P. M., Aaronson, N. K., Bottomley, A., de Graeff, A., Groenvold, M., ... & Quality of Life Cross-Cultural Meta-Analysis Group. (2009). A simulation study provided sample size guidance for differential item functioning (DIF) studies using short scales. Journal of clinical epidemiology, 62(3), 288-295. https://doi.org/10.1016/j.jclinepi.2008.06.003

Seidman, I., Yirmiya, N., Milshtein, S., Ebstein, R. P., & Levi, S. (2012). The Broad Autism Phenotype Questionnaire: Mothers versus fathers of children with an autism spectrum disorder. *Journal of Autism and Developmental Disorders, 42*(5), 837–846. https://doi.org/10.1007/s10803-011-1315-9

Sharma, Ujjwal; Bhushan, Braj (2018). The Hindi Adaptation and Standardization of BAPQ: A Brief Tool for Assessing Broad Autism Phenotype. Asian Journal of Psychiatry, S1876201818307494–. https://doi.org/10.1016/j.ajp.2018.10.008

Shi, L.-J., Ou, J.-J., Gong, J.-B., Wang, S.-H., Zhou, Y.-Y., Zhu, F.-R., … Luo, X.-R. (2015). Broad autism phenotype features of Chinese parents with autistic children and their associations with severity of social impairment in probands. BMC Psychiatry, 15(1). https://doi.org/10.1186/s12888-015-0568-9

Smith, A. B., Rush, R., Fallowfield, L. J., Velikova, G., & Sharpe, M. (2008). Rasch fit statistics and sample size considerations for polytomous data. BMC Medical Research Methodology, 8(1), 1-11. https://doi.org/10.1186/1471-2288-8-33

Stojković, I., Dučić, B., Kaljača, S., & Đorđević, M. (2018). Construction and psychometric evaluation of a short form of the Broad Autism Phenotype Questionnaire. *Psihologija, 51*(2), 243–258. https://doi.org/10.2298/PSI170619014S

Sucksmith, E., Roth, I., & Hoekstra, R. A. (2011). Autistic traits below the clinical threshold: Re-examining the broader autism phenotype in the 21st century. *Neuropsychology Review, 21*, 360–389. https://doi.org/10.1007/s11065-011-9183-9

Wainer, A. L., Ingersoll, B. R., & Hopwood, C. J. (2011). The Structure and Nature of the Broader Autism Phenotype in a Non-Clinical Sample. *Journal of Psychopathology and Behavioral Assessment, 33*, 459–469. https://doi.org/10.1007/s10862-011-9259-0

Wright, B. D., & Linacre, J. M. (1994). Reasonable mean-square fit values. Rasch Measurement Transactions, 8, 370–371 retrieve from https://www.rasch.org/rmt/contents.htm

Wright, B. D., & Masters, G. N. (1982). *Rating scale analysis*. MESA press.

Wright, B. & Stone, M. (1979). Best test design. MESA Press: Chicago, IL